PB81-182057

Evaluation of the RAM (Real-Time Air-Quality Model) Using the RAPS (Regional Air Pollution Study) Data Base

SRI International
Menlo Park, CA

Prepared for

Environmental Sciences Research Lab.
Research Triangle Park, NC

Mar 81

U.S. Department of Commerce
National Technical Information Service
NTIS.

TECHNICAL REPORT DATA
(Please read Instructions on the reverse before completing)

1. REPORT NO.	2.	3. RECIPIENT'S ACCESSION NO.
EPA-600/4-81-020		PB81 182057

4. TITLE AND SUBTITLE	5. REPORT DATE
EVALUATION OF THE RAM USING THE RAPS DATA BASE	March 1981
	6. PERFORMING ORGANIZATION CODE
7. AUTHOR(S)	8. PERFORMING ORGANIZATION REPORT NO.
Ronald E. Ruff	
9. PERFORMING ORGANIZATION NAME AND ADDRESS	10. PROGRAM ELEMENT NO.
SRI International 333 Ravenswood Avenue Menlo Park, California 94025	CDTA1D/03-0266 (FY-81)
	11. CONTRACT/GRANT NO.
	68-02-2770
12. SPONSORING AGENCY NAME AND ADDRESS	13. TYPE OF REPORT AND PERIOD COVERED
Environmental Sciences Research Laboratory, RTP, NC Office of Research and Development U.S. Environmental Protection Agency Research Triangle Park, North Carolina 27711	Final 8/77 - 4/79
	14. SPONSORING AGENCY CODE
	EPA/600/09

15. SUPPLEMENTARY NOTES

Previous Reports: EPA-600/4-80-013a, b, c, and d.

16. ABSTRACT

Application of statistical tests to the RAM (Real-Time Air-Quality Model) using the RAPS (Regional Air Pollution Study) data base is reported. Earlier work described the statistical tests in detail. Application of the statistical tests as described herein fulfilled two purposes: to quantify the overall performance of the RAM and to provide diagnostic information about the causes of poor model performance. Results of the diagnostic tests serve to identify specific weaknesses in the model formulation and some localized errors in the emission inventory.

Two microfiche accompany this report: 1976 STL SO2 RAM HOURLY DATA BASE, and 1976 STL SO2 RAM EVALUATION PROGRAM LISTINGS.

17. KEY WORDS AND DOCUMENT ANALYSIS

a. DESCRIPTORS	b. IDENTIFIERS/OPEN ENDED TERMS	c. COSATI Field/Group

18. DISTRIBUTION STATEMENT	19. SECURITY CLASS *(This Report)*	21. NO. OF PAGES
RELEASE TO PUBLIC	UNCLASSIFIED	
	20. SECURITY CLASS *(This page)*	22. PRICE
	UNCLASSIFIED	

EPA Form 2220-1 (9-73)

BEGIN FIRST LINE OF TEXT HERE
TYPING GUIDE SHEET
CENTER OF PAGE
TOP OF IMAGE AREA
DROPPED HEAD; BEGIN SECTIONS HERE
6-1/2"
9-1/8"
BEGIN LAST LINE OF TEXT
3/8"
BOTTOM OF IMAGE AREA; OUTSIDE DIMENSION FOR TABLES AND ILLUSTRATIONS
PAGE NUMBER
EPA-287 (Cin.)
(4-76)

DISCALIMER

This report has been reviewed by the Environmental Sciences Research Laboratory, U. S. Environmental Protection Agency, and approved for publication. Mention of trade names or commercial products does not constitute endorsement or recommendation for use.

BEGIN FIRST LINE OF TEXT HERE
TYPING GUIDE SHEET
CENTER OF PAGE
TOP OF IMAGE AREA
DROPPED HEAD. BEGIN SECTIONS HERE

ABSTRACT

Application of statistical tests to the RAM (Real-Time Air-Quality Model) using the RAPS (Regional Air Pollution Study) data base is reported. Earlier work described the statistical tests in detail. Application of the statistical tests as described herein fulfilled two purposes: to quantify the overall performance of the RAM and to provide diagnostic information about the causes of poor model performance. Results of the diagnostic tests serve to identify specific weaknesses in the model formulation and some localized errors in the emission inventory.

Two microfiche accompany this report: 1976 STL SO2 RAM HOURLY DATA BASE, and 1976 STL SO2 RAM EVALUATION PROGRAM LISTINGS.

6 1/2"
9 1/8"
BEGIN LAST LINE OF TEXT
BOTTOM OF IMAGE AREA OUTSIDE DIMENSION FOR TABLES AND ILLUSTRATIONS
3/8"
PAGE NUMBER
EPA-287 (Cln-1 14-76)

BEGIN FIRST LINE OF TEXT HERE
TYPING GUIDE SHEET
CENTER OF PAGE
TOP OF IMAGE AREA
DROPPED HEAD: BEGIN SECTIONS HERE

CONTENTS

Abstract . . . . . **iii**
Figures . . . . . **vii**
Tables . . . . . **x**
Acknowledgments . . . . . **xi**

1. Overview 1
   - Background 1
   - Conclusions 1
2. The RAMS/RAPS data base 4
   - The RAM 4
   - The RAPS data base 4
   - Description of the RAMS/RAPS data base 4
3. Analytical Method 7
   - Distinction between method objectives 7
   - Final evaluation statistics 7
     - Objective loss functions 8
     - Subjective loss functions 9
     - Loss function for the location of maximum concentration 10
   - Diagnostic methods 11
4. The Model Evaluation Process 13
   - General procedure 13
   - Diagnosing the causes of poor model performance . . 13
   - Order of applying diagnostic tests 14
5. The RAM Evaluation 17
   - General discussion 17
   - Accuracy score results and characteristics 17
     - Results for the hourly average at each station 17
     - Averaging time characteristics 19
     - Application of the mean-square error to two versions of RAM 19
   - Application of diagnostic methods 22
     - Detection of bias 22
     - Detection of randomness. 24
     - Selection of specific problem areas 27
     - Diagnosing the causes of bias at Site 101 . . 27
     - Diagnosing the causes of bias at Site 104 . . 30
     - Comparison of results from other sites with results from Sites 101 and 104 40
   - Future analysis 40

BEGIN LAST LINE OF TEXT
3/8"
DIMENSION FOR TABLES AND ILLUSTRATIONS
PAGE NUMBER
EPA-287 (Cin.) (4-76)

Preceding page blank

Top of 8½ x 11 paper

CLASSIFICATION

PB81-182057

EPA-600/4-81-020
March 1981

EVALUATION OF THE RAM USING THE RAPS DATA BASE

by

Ronald E. Ruff
Atmospheric Science Center
SRI International
Menlo Park, California 94025

Contract 68-02-2770

Project Officer

John S. Irwin
Meteorology and Assessment Division
Environmental Sciences Research Laboratory
Research Triangle Park, North Carolina 27711

This study was conducted
in cooperation with
SRI International
Menlo Park, California 94025

ENVIRONMENTAL SCIENCES RESEARCH LABORATORY
OFFICE OF RESEARCH AND DEVELOPMENT
U.S. ENVIRONMENTAL PROTECTION AGENCY
RESEARCH TRIANGLE PARK, NORTH CAROLINA 27711

REPRODUCED BY
NATIONAL TECHNICAL
INFORMATION SERVICE
U.S. DEPARTMENT OF COMMERCE
SPRINGFIELD, VA 22161

Top of 8½ x 11 paper
EPA ORD REPORT

6. Assessment of Alternative Statistics 43
Relative merit among tests 43
Alternative final evaluation statistics 45
Other research 45
Using attributes of the frequency distributions as a criteria 45
7. Software 50
Overview 50
File descriptions 50

References 54
Appendices

A. Frequency distribution of RAPS hourly meteorological parameters 56
B. Logarithm-frequency graphs of observed and predicted concentrations 61
C. Scatterplots of hourly predicted versus observed SO_2 concentrations for selected RAMS monitoring sites 69
D. Irwin's results for Site 104 73

Bottom of 8½ x 11 paper

Top of 8½ x 11 paper

FIGURES

Number		Page
1	Layout of the RAPS SO_2 monitoring stations showing annually averaged observed and (RAM) predicted concentrations ($\mu g/m^3$)	5
2	Flow chart of diagnostic steps	15
3	Effects of averaging time on mean square error	20
4	Effects of averaging time on accuracy of predicting location of maximum concentration	20
5	Logarithm frequency distributions for selected sites	23
6	Interstation error correlation test results	25
7	Scatterplot for three-hour averaged predicted values versus observed concentrations	26
8	Multiple regression summary table for a wind sector from 279° to 288°--Station 101	28
9	Scattergram of error residual concentrations versus atmospheric stability class (Station 101; wind directions from 279° to 288°).	29
10	Discriminant group summaries for agreement, underprediction, and overprediction using 300 $\mu g/m^3$ threshold	32
11	Discriminant group summaries for agreement, underprediction, and overprediction using 600 $\mu g/m^3$ threshold	33
12	Prediction accuracy of discriminant function--600 $\mu g/m^3$ threshold	34
13	Time sequence of high observed concentration at Site 104 during December 1976	35
14	Partial CROSSTABS output for Site 104	36
15	Point source locations in relation to RAPS monitoring stations with a background of area source grids	39

Bottom of 8½ x 11 paper

Top of 8½ x 11 paper
EPA ORD REPORT

Number		Page
16	Autocorrelation of error residual versus time lags (in hours)	44
17	Frequency distribution by class on a linear scale for Station 101	47
18	Computer tables of three-hour averaged ten highest observed and predicted concentrations--all sites	49
19	Example of data retrieval session using RAMDATARET	53
A-1	Frequency distribution of atmospheric stability classes for 1976	57
A-2	Distribution of wind speed for 1976	57
A-3	Frequency distribution of wind direction for 1976	58
A-4	Frequency distribution of temperature for 1976	58
A-5	Frequency distribution of point source emissions for 1976	59
A-6	Frequency distribution of area source emissions for 1976	59
A-7	Frequency distribution of mixing height for 1976	60
B-1	Observed and predicted concentrations for Station 101	62
B-2	Observed and predicted concentrations for Station 103	62
B-3	Observed and predicted concentrations for Station 104	63
B-4	Observed and predicted concentrations for Station 105	63
B-5	Observed and predicted concentrations for Station 106	64
B-6	Observed and predicted concentrations for Station 108	64
B-7	Observed and predicted concentrations for Station 113	65
B-8	Observed and predicted concentrations for Station 114	65
B-9	Observed and predicted concentrations for Station 115	66
B-10	Observed and predicted concentrations for Station 116	66
B-11	Observed and predicted concentrations for Station 120	67
B-12	Observed and predicted concentrations for Station 121	67

Bottom of 8½ x 11 paper

Top of 8½ x 11 paper

Number		Page
B-13	Observed and predicted concentrations for Station 122	68
B-14	Composite observed and predicted concentrations for all stations	68
C-1	Scatterplot of all 1976 hourly SO_2 averages for Site 101 . . .	70
C-2	Scatterplot of all 1976 hourly SO_2 averages for Site 103 . . .	70
C-3	Scatterplot of all 1976 hourly SO_2 averages for Site 104 . . .	71
C-4	Scatterplot of all 1976 hourly SO_2 averages for Site 106 . . .	71
C-5	Scatterplot of all 1976 hourly SO_2 averages for Site 116 . . .	72
C-6	Scatterplot of all 1976 hourly SO_2 averages for Site 122 . . .	72
D-1	Distribution of Site 104 residuals by wind direction	78
D-2	Distribution of Site 104 residuals by wind speed	79
D-3	Distribution of Site 104 residuals by mixing height.	80
D-4	Distribution of Site 104 residuals by Pasquill Gifford stability class.	81
D-5	Distribution of Site 104 residuals by time of day.	82
D-6	Distribution of overprediction and underprediction for 1976. .	83

Bottom of 8½ x 11 paper

Top of 8½ x 11 paper
EPA ORD REPORT

TABLES

Number		Page
1	Accuracy score results	18
2	Accuracy score (mean-squared error) comparison between two wind-averaging schemes	21
3	Site 104 underprediction of 600 $\mu g/m^3$ and greater by wind direction	37
4	Irwin's preliminary conclusions for Site 104 (1976)	38
5	Summary of multiple regression statistics for selected sites	41
6	Value of chi-square as a function of site for hourly averages	48
7	Format for the 1976 hourly data base	51
D-1	Summary of residuals for Jan-Feb-Mar 1976	74
D-2	Summary of residuals for Apr-May-Jun 1976	75
D-3	Summary of residuals for Jul-Aug-Sep 1976	76
D-4	Summary of residuals for Oct-Nov-Dec 1976	77

Bottom of 8½ x 11 paper

Top of 8½ x 11 paper

EPA ORD REPORT

ACKNOWLEDGMENTS

I would like to express my gratitude to the people who gave support and assistance during the course of this project.

First of all, I wish to acknowledge the extraordinary efforts of my colleague, Dr. Rodney H. Allen of Comp-Aid, Incorporated. Dr. Allen adapted all the SRI software to the EPA UNIVAC 1110 system, finding and correcting many problems in the process. Additionally, he wrote many special data-handling programs and made countless improvements on existing programs. His dedication and technical expertise were essential to the successful conduct of this work.

Two of my colleagues at EPA also made significant contributions to the work reported herein. The project officer, Mr. John S. Irwin, provided the necessary technical coordination and in-depth review. Furthermore, he made substantial technical contributions, some of which are singled out in the text. His assistance and advice were invaluable in solving many problems that developed during the project. Another EPA colleague, Ms. Joan H. Novak, was instrumental in the assembly of the RAM and RAPS data base and answering many technical questions that arose about them. I am grateful for her timely and accurate efforts.

Finally, the support of my fellow staff members at SRI is appreciated. Dr. Harold S. Javitz deserves recognition for providing timely consultation on the interpretation of results. Our secretaries--Ms. Dorothy Sevela, Ms. Barbara Wildman, and Ms. Shirley Webster--deserve credit for typing and formatting the draft manuscript so that it was easy to review and transform into the final version.

Bottom of 8½ x 11 paper

SECTION 1

OVERVIEW

BACKGROUND

Until recently, the ability to provide objective evaluations of air-quality models has suffered for two reasons:

- The scarcity of adequate test data bases
- Inadequate statistical procedures.

The first problem has been alleviated somewhat through the acquisition of the extensive Regional Air Pollution Study (RAPS) data base. The primary objective of the RAPS was to provide a data base useful in evaluating, upgrading, and developing air quality models.

With the availability of the comprehensive RAPS data base in mind, EPA retained SRI International to develop a statistical package that would overcome the limitations of currently used statistical procedures. While EPA sought a generally applicable statistical package, they recognized the need to test the package, and for this purpose chose the Real-Time Air-Quality Model (RAM) along with the RAPS data base.

The statistical package was developed and tested. A description of this effort is given in a four-volume report (Ruff, 1980a; Javitz and Ruff, 1980; Ruff et al., 1980; Ruff 1980b). This volume describes results of applying the statistical package to the RAM evaluation.

CONCLUSIONS

The tests were adequate in describing the RAM performance and aiding in the diagnosis of several theorized causes for isolated cases of poor performance. This twofold purpose of the statistical tests--validation and diagnosis--is the critical point here. We feel that the statistical tests described herein have been demonstrated to satisfy both objectives. This is not to say that the specific protocol and tests presented are recommended as the nationwide criteria for model evaluation. Rather, we hope that this work will contribute to such criteria development, along with work from other investigators.

Applying the tests to the RAM produced an enormous amount of data. Some results were quite useful in helping us develop theories concerning some aspect of the model's performance; other results were not easy to interpret

and some of the tests were not as useful as we had hoped. However, we do not feel that any of the tests are inappropriate based on this one application.

We concluded a number of things about the RAM performance. Basically, it does not perform well on a case-by-case basis when comparing observed and predicted results at the same place and the same time. Typically, the correlation coefficient is below 0.25 when comparing the agreement between concurrent hourly averaged observed and predicted concentrations. This poor correlation is examined in various ways throughout this text and is graphically demonstrated with scattergrams.

The RAM performs fairly well when comparing the frequency distribution of observed concentrations with the frequency distribution of predicted concentration on a site-by-site and overall basis. Our results do not agree with those displayed by Guldberg and Kern (1978) for their analysis of 24-hour concentrations in Michigan City and Indianapolis, Indiana. Their results showed RAM overpredicted by factors of two to four, depending on the site. Our analysis using the RAPS data base indicated much better agreement. For concentrations greater than 20 $\mu g/m^3$, the agreement was generally within 50 percent. Analysis of the annual averages for each site leads us to conclude that there is no general tendency of RAM to overpredict or underpredict. For many sites, the annually averaged predictions agree with observations within experimental error. At two sites (106 and 116), the annual averages perfectly agree. At those sites with the highest misprediction, explanations presented in the next paragraph attribute most of the misprediction to problems with the model input parameters rather than the model formulation.

Three different potential causes of poor model performance were identified. These causes and our recommendations for remedies appear below:

- The model tends to overpredict at locations near the center of the city. In part, this can be attributed to the manner by which a large number of small sources are represented by one area source with a single release height. One remedy for this problem would be to break down further the individual sources within the areas with larger emissions: either representing more of them as point sources, using smaller area boundaries, or using more than one height of emission per area source. Another possible remedy might be to increase the initial value of the vertical stability parameter.
- One site (104) exhibited severe underprediction that was influenced by a fairly small number of cases (about 100), mostly occurring within a four-week time span. We believe this problem is attributable to an uncatalogued emission, possibly from an already catalogued source in the vicinity of the site. SRI recommends that attempts be made to verify this theory and, if correct, quantify the emissions. If this cannot be done, SRI recommends that the contributing cases be dropped from the evaluation data base. (It is also possible that the emission is catalogued and that the underpredictions result from some specialized local effect.)

- All sites exhibit extensive random behavior when comparing concentrations for the same place and time. Much of this randomness is believed to be attributable to the misrepresentation of the wind vector (particularly direction). This results from the model simplification of assigning the same wind vector to entire area. This simplification is a limitation in RAM; the solution requires extensive modification using more spatial or temporal resolution in the wind field. However, this requires more wind measurements than available in most urban areas.

SECTION 2

THE RAMS/RAPS DATA BASE

THE RAM

The RAM (Turner and Novak, 1978) was developed by EPA several years ago as a tool for calculating SO_2 surface concentrations of 1- and 24-hour averaging times. These can be used to form estimates of SO_2 concentrations for 3-hour, monthly, seasonal, and annual averaging times. This Gaussian-plume multiple-source (point and area) model requires hourly input data for emissions and meteorological parameters from a single site. Hence, the model output consists of an SO_2 concentration prediction for each hour.

THE RAPS DATA BASE

The RAPS was conducted in St. Louis because the geographic area already had a fairly good historic air-quality data base and was one of the least complicated urban areas to model. (St. Louis has fairly level terrain and is free of pollutant intrusion from other large urban areas.) The heart of the RAPS is the 25-station air-monitoring network that collects data on the criteria pollutants (including SO_2) and meteorological parameters (wind, ambient temperature, dew point temperature, and temperature gradient). The geographical layout of the 13 network stations used in our analysis is shown in Figure 1. In addition to the network, numerous special measurements were made at the monitoring sites, other surface locations, and aloft. Hence, the data base is rich not only in the SO_2 measurements needed to evaluate model performance, but also in the emission and meteorological measurements needed to understand the limitations of the RAM input parameters. Other documents (e.g., Schiermeier, 1978) provide comprehensive documentation on RAPS.

DESCRIPTION OF THE RAM/RAPS DATA BASE

The RAM was executed for each hour of the year 1976 using meteorological and emission input parameters from the RAPS data base. The wind parameter was converted to a single wind direction and speed by interpolating wind measurements from the 25 monitoring stations. In all, 13 receptor locations were used in the analysis. The model evaluation data base for each hour includes calendar date, time, wind speed, wind direction, atmospheric stability class, mixing height, total point source emissions, total area source emissions, and thirteen pairs of observed and predicted SO_2 concentrations. From the hourly data base of predictions and observations, data sets were constructed for longer averaging times--3-hourly, 8-hourly, daily, weekly, monthly, 3-monthly, and yearly. Frequency distributions are given in

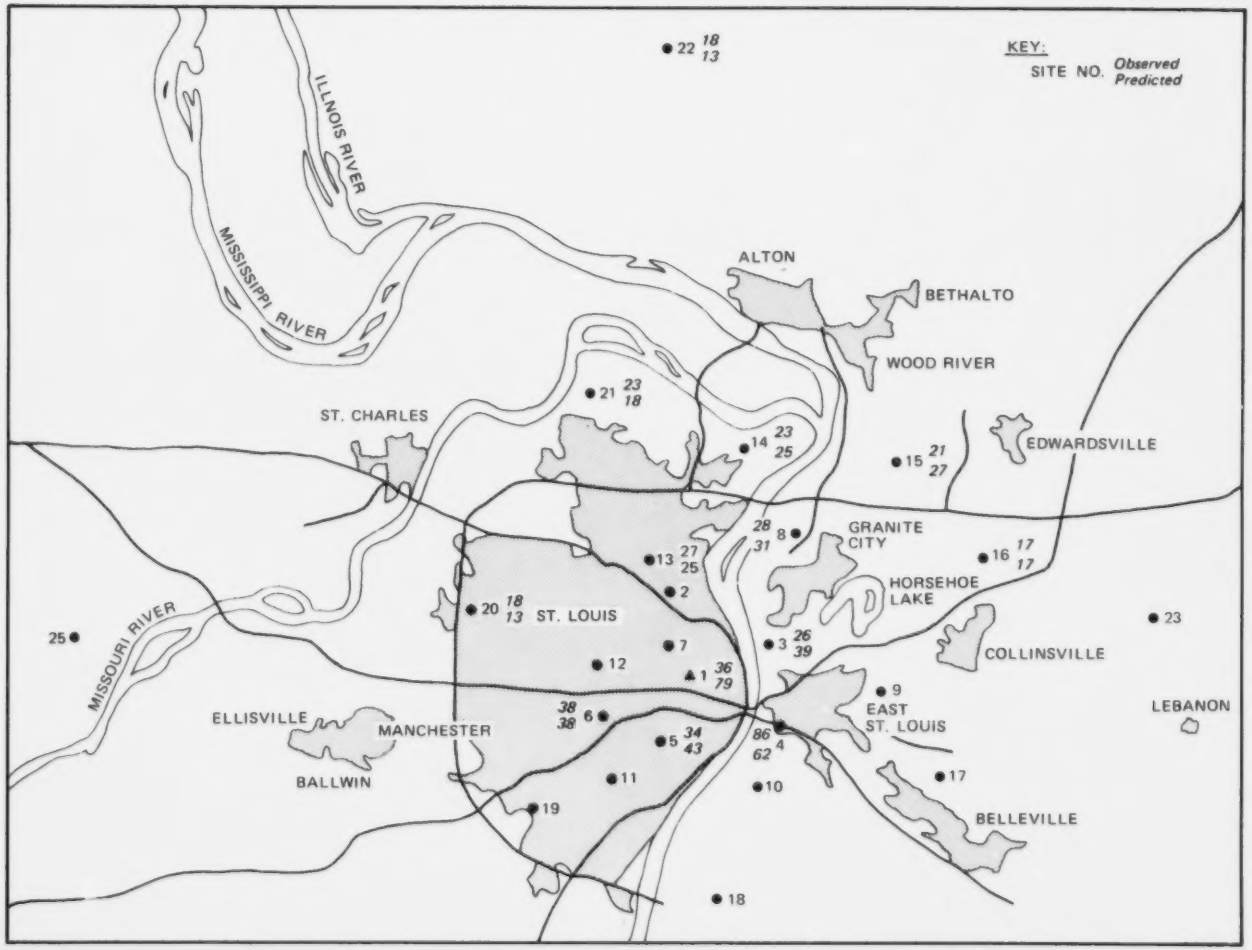


Figure 1. Layout of RAPS SO_2 monitoring stations showing annually averaged observed and (RAM) predicted concentrations ($\mu g/m^3$).

Add 100 to site designators to get standard RAPS site designator.

Top of 8½ x 11 paper

Appendix A for the meteorological model input parameters, and aggregated area and point source emissions.

The observed SO_2 concentrations underwent extensive quality assurance tests (Klein and Littman, 1978; Hribar, 1976). However SO_2 measurement errors are unavoidable and influence the comparisons that follow. In particular, note that concentrations below 6 $\mu g/m^3$ were not believed to be accurately represented, hence, the minimum-reported observation was always rounded up to 6 $\mu g/m^3$ and reported as measured for concentrations higher than 6 $\mu g/m^3$.

Predicted concentrations are, of course, subject to errors in the model input data and in the model formulation. The evaluation, for instance, assumes all the significant SO_2 emission sources have been accounted for. Background sources are assumed to be negligible.

Bottom of 8½ x 11 paper

Top of 8½ x 11 paper

EPA ORD REPORT

SECTION 3

ANALYTICAL METHOD

DISTINCTION BETWEEN METHOD OBJECTIVES

There are two distinct objectives in applying the various methods:

- Final evaluation--An accuracy score or figure of merit that forms the basis of comparison among competing models or different versions of the same model. This score summarizes model performance.
- Diagnostic (intermediate) evaluation--Statistical and graphical methods are used to provide insight as to why a model performs poorly (or well).

Certain types of methods are better suited to diagnostic evaluation and others are better suited to final evaluations. There are no hard and fast lines, however, and a statistic may be quite useful in satisfying both objectives.

FINAL EVALUATION STATISTICS

A set of final evaluation statistics (termed "accuracy scores") were developed for a variety of applications. The accuracy score offers the user several tests to assess either objectively or subjectively the "loss" resulting from differences in predicted concentrations (PC) and observed concentrations (OC). Each test is based on a quantified loss function, which measures the loss associated with an error in prediction at a particular time period. The accuracy score (denoted A) is the average loss for the prediction model computed over the n time periods in the data base. Consequently, the accuracy score resembles a golf score: the lower, the better.

We considered eight loss functions, which naturally are associated with eight different accuracy scores. The first five rely on (essentially) prespecified loss functions, which are easily interpretable and have therefore been called objective. The next two require that the user specify loss indices associated with various types of misprediction and have therefore been called subjective. The eighth loss function provides an index of the spatial accuracy of the model in predicting the location of the maximum observed concentration.

The importance of correctly specifying the loss function cannot be overemphasized. The loss function embodies our conception of how poorly or well the model fits and the relative seriousness of different departures from perfect fit. For example, one who is concerned only about evaluating a model

Bottom of 8½ x 11 paper

will choose a different loss function than one who is only concerned with the ability to predict actual alert-level pollution (i.e., exceedences).

We now describe eight loss functions and accuracy scores--five objective and three subjective.

Objective Loss Functions

Average Absolute Error--

$$A = \frac{1}{n} \sum_{i=1}^{n} |OC_i - PC_i| \quad ,$$

where

A is the accuracy score

OC_i is the observed concentration ($\mu g/m^2$) for time period i

PC_i is the predicted concentration ($\mu g/m^3$) for time period i

n is the number of OC_i and PC_i pairs.

With this absolute error loss function, underprediction and overprediction are judged to be equally undesirable. Furthermore, a misprediction of 2 ppm is twice as undesirable as a misprediction of 1 ppm.

Average-Squared Error--

$$A = \frac{1}{n} \sum_{i=1}^{n} (OC_i - PC_i)^2 \quad .$$

With this squared error loss function, underprediction and overprediction also are judged to be equally undesirable. A misprediction of 2 ppm is four times as undesirable as a misprediction of 1 ppm. This type of loss function is used in many statistical tests.

Percent Incorrect Predictions with Absolute Error Threshold (ε)--

$$A = \frac{1}{n} \sum_{i=1}^{n} \begin{cases} 0 & \text{if } |OC_i - PC_i| \leq \varepsilon \\ 1 & \text{if } |OC_i - PC_i| > \varepsilon \end{cases} \quad .$$

With this loss function, only prediction errors that exceed a certain absolute threshold are counted against the prediction model. Underprediction and overprediction are judged to be equally undesirable. Whenever the misprediction

Top of 8½ x 11 paper

exceeds the threshold, its magnitude is ignored. Thus, moderate and huge misprediction errors count equally. (The error threshold must be user-supplied.)

Percent Incorrect Prediction with Constant Percentage Error Threshold (ε)--

$$A = \frac{1}{n}\sum_{i=1}^{n}\begin{cases} 0 \text{ if } \left|\dfrac{OC_i - PC_i}{OC_i}\right| \leq \varepsilon \\ 1 \text{ if } \left|\dfrac{OC_i - PC_i}{OC_i}\right| > \varepsilon \end{cases} .$$

This loss function is similar to the preceding, except that the threshold criterion is percentage error. (In practice, the case where OC = 0 is not allowed.)

Symmetric High-Low Loss Function (Cutoff)--

$$A = \frac{1}{n}\sum_{i=1}^{n}\begin{cases} 0 \text{ if } \begin{matrix} PC_i \text{ and } OC_i \leq \text{cutoff} \\ \text{or} \\ PC_i \text{ and } OC_i > \text{cutoff} \end{matrix} \\ 1 \text{ if otherwise} \end{cases} .$$

This loss function is particularly appropriate when an alert level cutoff exists. Here A is the percentage of false alerts and failures to predict actual pollutant alert levels. The magnitude of the false alert is ignored, as is the magnitude of the underprediction when an actual pollutant alert condition is not predicted.

Subjective Loss Functions

Asymmetric High-Low Loss Function (Cutoff)--

$$A = \frac{1}{n}\sum_{i=1}^{n}\begin{cases} 0 \text{ if or } \begin{matrix} PC_i \text{ and } OC_i \leq \text{cutoff} \\ PC_i \text{ and } OC_i > \text{cutoff} \end{matrix} \\ \ell_1 \text{ if } PC_i > \text{cutoff and } OC_i \leq \text{cutoff} \\ \ell_2 \text{ if } PC_i \leq \text{cutoff and } OC_i > \text{cutoff} \end{cases} .$$

Here, ℓ_1, ℓ_2, and the cutoff value are user-specified loss quantities. This loss function is similar to the preceding one, except that underprediction and overprediction are weighted differently.

Page Number

Bottom of 8½ x 11 paper

User-Supplied Loss Matrix--

The predicted and observed concentrations are divided into categories whose increments are about 20 percent of the applicable standard. The user enters the loss numbers in the matrix. An example of the loss table is shown below:

PC Categories (j) \ OC Categories (k)	1	2	3	4	5
1	0	1	2	3	4
2	1	0	1	2	3
3	2	1	0	1	2
4	3	2	1	0	1
5	4	3	2	1	0

Let ℓ_{jk} be the number in the jth row and the kth column. The accuracy score is

$$A = \frac{1}{n} \sum_{i=1}^{n} \ell_{jk} \quad ,$$

where PC_i is in the jth category and OC_i is in the kth category.

The loss function, ℓ_{jk}, is extremely general, since the number of cells and the loss numbers in the cells are at the user's discretion. The sample loss matrix approximates the average absolute error loss function; other loss functions could be similarly approximated.

Loss Function for the Location of Maximum Concentration

The previous concepts of the accuracy score can be slightly modified to provide an index of the spatial accuracy of the prediction model in forecasting the location of the maximum-concentration reading. This accuracy score can be defined in the following fashion:

Step 1: For the ith time period, identify the monitoring locations, k(i) and k'(i), where the highest observed and predicted concentrations occur.

Step 2: From the symmetric m × m matrix, consisting of the distances between monitoring stations, select the distance $d_{k(i),k'(i)}$ between the locations for highest observed and predicted concentrations.

Step 3: Repeat Steps 1 and 2 for all samples and then compute the accuracy score from the formula

$$A = \frac{1}{n} \sum_{i=1}^{n} d_{k(i),k'(i)} \quad .$$

Many other types of objective or subjective loss indices could be used instead of geographical distance. For instance, the loss index could be constructed to take into account topographical features that might cause two geographically close locations to differ substantially in pollutant levels.

DIAGNOSTIC METHODS

A number of statistical tests were considered for intermediate evaluation. (However, there are special cases where these tests can also be used as a final evaluation tool.) The following tests and displays are recommended and included as part of the software package:

- Scatterplot
- Linear regression with confidence and prediction bands
- Pearson's correlation coefficient
- Chi-square test and frequency distribution displays
- Interstation error correlation
- Analysis of residual time series
- Multiple regressions of error residuals
- Analysis of discriminant functions.

In the software package, as discussed later, some of the above tests were combined to make computer execution more efficient. In the following paragraphs, the tests are considered separately.

A simple scatterplot, also known as a scattergram, is one of the most useful evaluation tools because it displays a case-by-case comparison of all data (i.e., observed versus predicted concentrations). A regression analysis can be added to compute and display (on the scatterplot) a least-squares fit of the data. Confidence and probability bands can be calculated and added to the display to demonstrate the statistical significance of the regression (least-squares) line and the probability of a new test point falling outside the bands, respectively. The Pearson's correlation coefficient is calculated as a measure of how linear the comparison is.

The chi-square test and accompanying frequency distribution display are a measure of how well the predicted frequency distribution agrees with the observed frequency distribution. This test is also a useful final evaluation statistic when the distribution itself is important and not the point-by-point comparison.

The interstation error correlation test examines the correlation between station errors. (Error is the difference between observed and predicted concentrations). This test is useful for identifying geographical (or

Top of 8½ x 11 paper

spatial) bias in the model comparison process. The analysis of residual time series test examines the errors for any cyclical behavior. It determines whether errors occur on a periodic basis and what the time period is. Correlation of the error (residuals) with the model input parameters is computed by the multiple regression of error residuals test. The primary purpose of this test is to diagnose which of the input parameters most influence the model's poor performance. Discriminant analysis allows one to analyze meteorological and emission conditions that correspond to some user-defined criteria, such as large overpredictions or underpredictions. This form of analysis is particularly useful in identifying whether a poor model performance criterion is linked to a specific set of model input conditions.

Bottom of 8½ x 11 paper

Top of 8½ x 11 paper

SECTION 4

THE MODEL EVALUATION PROCESS

GENERAL PROCEDURE

The general method of applying the final evaluation and the diagnostic tests remains essentially the same, regardless of the specific air-quality model being tested. There are two primary objectives in evaluating models

- Comparison among competing models
- Improvement of a particular model.

The first objective requires only the computation of the final evaluation statistics for each competing model; the second objective benefits from both the final evaluation and diagnostic methods.

The basic order of events in evaluating and improving the RAM (or any model) is to:

- Apply the final-evaluation statistics
- Apply the diagnostic methods
- Interpret the results
- Run tests on subgroups
- Modify the model (RAM)
- Apply the final-evaluation statistics to the modified model
- Repeat the above process until the desired improvement is achieved.

The second through fourth steps are repeated in an iterative fashion. As test results are analyzed, the evaluator tries another test and so forth until full value has been garnered from the diagnostic tests. These tests are quite inexpensive to run in comparison to expenses associated with modifying the model and regenerating a test data base.

DIAGNOSING THE CAUSES OF POOR MODEL PERFORMANCE

There is some redundancy in our package, which means that conclusions about RAM performance may be arrived at in a number of different manners. When evaluating poor performance, we are looking for one or more types of poor-performance indicators. For each such indicator, one or more statistical

Bottom of 8½ x 11 paper

tests will reveal something about the RAM performance. These indicators* of poor performance include:

- General bias occurs when the differences in observed and predicted concentrations are mostly of the same magnitude and sign. Several of the tests will detect this condition.
- Spatial bias is similar to general bias, differing in that the spatial patterns are examined. The same tests for general bias are applicable here, but they are applied to each site. In addition, the interstation error correlation test is useful in addressing error residuals and then spatial pattern.
- Temporal (cyclic) bias is detectable by the residual time series test. Temporal bias is indicated when the errors in misprediction are periodic.
- Randomness is indicated when poor agreement exists (between observed and predicted concentrations) and there is no apparent trend. This is best displayed in the bivariate regression and correlation program: visually with the scatterplot and numerically with a low Pearson's correlation coefficient.
- Input-parameter dependence exists when poor performance can be traced to one or more of the input parameters. The discriminant analysis and multiple regression of error residuals tests are useful for this purpose. Other tests, applied to subsets of the data base, can also be effective in detecting input parameter dependence.

The tests are able to detect all of the above indicators of poor model performance. However, as will be shown in Section 5, interpretation of the results leading to meaningful conclusions relies on a certain amount of subjective investigation (i.e., detective work).

ORDER OF APPLYING DIAGNOSTIC TESTS

The exact order of applying diagnostic tests is one of personal preference. In this section, we recommend a general procedure that could be followed. However, there are many variations that can lead to a successful diagnosis.

As shown in Figure 2, we recommend first applying the tests that detect bias. Once it has been determined that bias does not exist (or has been eliminated), we then change our focus to randomness. Basically, it is our view that bias is much easier to correct than randomness. Also, a model without bias is useful even in the presence of randomness. (This does not imply that a model with bias but no randomness is not useful.)

---

*Terms applied (e.g., randomness) are not designed to reflect standard statistical terminology; it is our intent for the terms to be readily relatable to the visual displays presented herein by a reader with only minimal training in statistics.

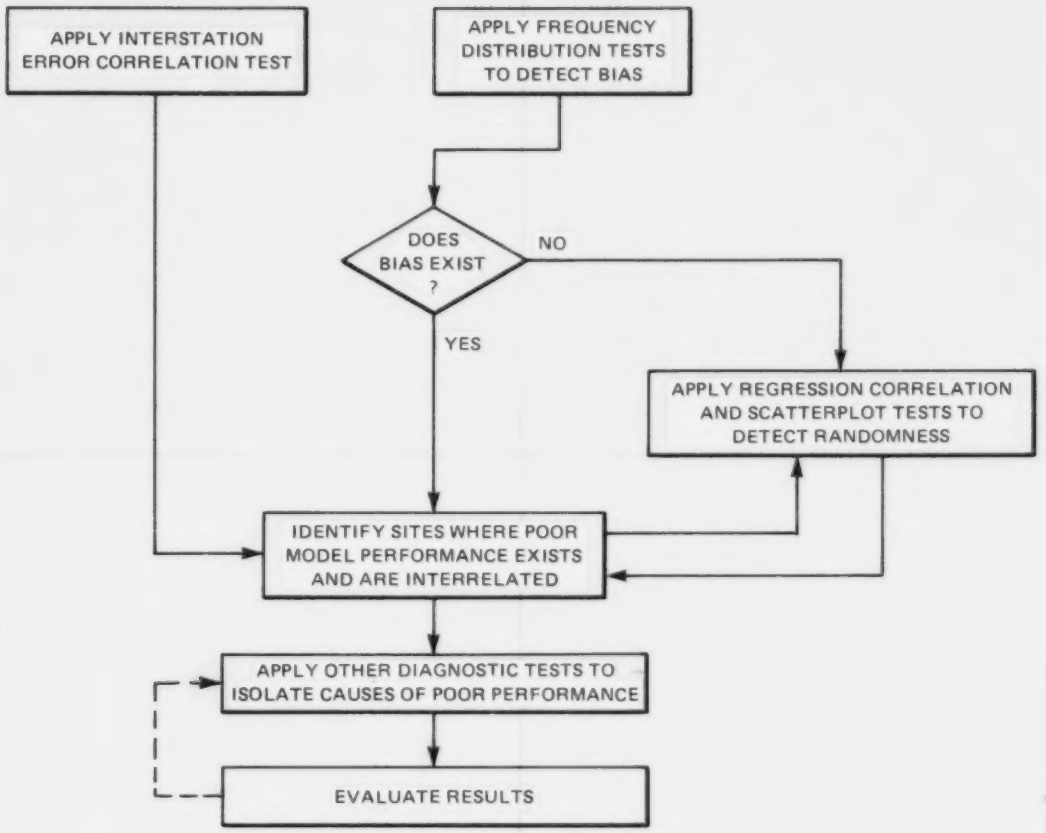

APPLY INTERSTATION ERROR CORRELATION TEST
APPLY FREQUENCY DISTRIBUTION TESTS TO DETECT BIAS
DOES BIAS EXIST ?
NO
YES
APPLY REGRESSION CORRELATION AND SCATTERPLOT TESTS TO DETECT RANDOMNESS
IDENTIFY SITES WHERE POOR MODEL PERFORMANCE EXISTS AND ARE INTERRELATED
APPLY OTHER DIAGNOSTIC TESTS TO ISOLATE CAUSES OF POOR PERFORMANCE
EVALUATE RESULTS

Figure 2. Flow chart of diagnostic steps.

Three tests are useful in quantifying and displaying bias: the chi-square test, frequency distributions on logarithmic axes, and display of annual averages on a geographical map. Application of all these tests provides some redundant information, but each also has unique features. For instance, frequency distributions on a logarithmic axis enable one to identify differences in extreme values (i.e., high concentrations). The chi-square test also displays somewhat less detailed frequency distributions, but at a much lower computer expense than the logarithmic displays. Additionally, the chi-square statistic computes an objective measure of how well the observed and predicted distributions agree. The geographical display provides a visual reference for assessing spatial bias. Through application of the above tests, we can determine whether we have general or spatial bias.

Having established the presence of bias, results from the interstation error correlation help us to identify whether the causes of bias are related from station to station. This helps identify which stations should be the target for further diagnostic tests, possibly including tests for randomness. For instance, if a bias is larger at some stations than others and the

interstation error correlation shows high correlations among certain of these stations, we may want to apply other diagnostic tests to these stations. The rationale is simple: If we identify the cause of the poor performance, one potential modification could improve performance at all sites where the bias is related.

In the absence of bias, we feel that a similar rationale holds for diagnosing the causes of randomness. As stated before, the regression and correlation analyses with scatterplot displays is a test for randomness. Specifically, we first identify where the payoff is potentially the greatest before applying further diagnostic tests. After identifying sites to focus on, we then apply the remaining tests--multiple regression of error residuals, discriminate analysis, and the (error) residual time series--on a selective basis, usually working on data from one site at a time. As results are evaluated, other tests are run until a cause of poor performance is identified. In this iterative process, results from one test are the basis for selection of another test.

At some point, we attempt to modify the model such that significant improvement results. Next, we apply the final evaluation statistics to compare performance between the new and old version of the model. We then repeat the diagnostic process with the goal of further improvements.

SECTION 5

THE RAM EVALUATION

GENERAL DISCUSSION

In this section, we present an example of properties of the RAM as quantified by selected final evaluation and diagnostic tests. The accuracy scores are, of course, the primary final evaluation statistic. However, we sometimes cast judgement on how well a model is performing based on one or the other of our diagnostic tests. In this context, bias and randomness are treated somewhat independently.

Lack of bias would indicate that over the time period of the data base (i.e., 1976), the model accurately represented all relevant meteorological and emission conditions. We acknowledge that, on an hour-to-hour basis, observed and predicted concentrations may not correspond at all (i.e., randomness). Lack of bias in the presence of randomness makes a model quite useful for some purposes (e.g., planning) and not for others (e.g., episode prediction). Hence, accuracy scores alone may not give the potential user all the "final evaluation" information needed to adjudge the overall utility of the model. Accordingly, comparisons of observed and predicted concentration frequency distributions (e.g., graphs or chi-squared tests) can be used as a final evaluation test in concert with the accuracy scores. However, we do emphasize that the primary goal of model improvement is the lowest possible accuracy score (i.e., agreement between observed and predicted concentrations on a case-by-case basis.)

ACCURACY SCORE RESULTS AND CHARACTERISTICS

Results for the Hourly Average at Each Station

The accuracy score tests were applied to the RAM/RAPS hourly data base for each station. The results are given in Table 1 for each of the first seven tests. (Results for Test 8--Maximum Concentration Location--are discussed later.) In comparing the results among tests, it is seen that scores in general are highest for Site 101 or Site 104, depending on the specific test. Additionally, we see that the scores are highest for sites closest to the center of the St. Louis urban area and become progressively lower as the distance from the center increases. (Note that the furthest outlying sites--116, 120, 121, and 122--have the lowest scores.)

One exception to the above generality is Test 4 where the threshold, E2, was set to 0.2. In this case, the highest possible score is unity. A score of unity would tell us that the observed and predicted concentrations always

TABLE 1. ACCURACY SCORE RESULTS

Site	Test 1 Mean absolute error	Test 2 Mean squared error	Test 3 Absolute error threshold ϵ_1 = 80 µg/m^3	Test 4 Percentage error threshold ϵ_2 = 0.2(20%)	Test 4 Percentage error threshold ϵ_2 = 2.0(200%)	Test 5 Symmetric hi-lo loss function ϵ_3 = 80 µg/m^3	Test 6 Asymmetric loss function ϵ_4 = 80 µg/m^3 ℓ_1=2, ℓ_2=1	Test 7 User-supplied loss matrix C_{max} = 1000 µg/m^3
101	71.1	12700	0.272	0.915	0.472	0.338	0.617	0.145
103	38.6	5170	0.141	0.892	0.270	0.193	0.337	0.050
104	99.2	7060	0.231	0.898	0.316	0.243	0.376	0.268
105	40.8	5180	0.144	0.881	0.228	0.192	0.299	0.049
106	44.0	7600	0.157	0.875	0.250	0.192	0.278	0.069
108	32.3	3510	0.115	0.879	0.195	0.162	0.261	0.030
113	28.3	2700	0.109	0.931	0.138	0.136	0.208	0.023
114	30.2	3760	0.101	0.936	0.193	0.133	0.204	0.033
115	28.5	3440	0.101	0.934	0.170	0.115	0.188	0.037
116	21.1	2220	0.055	0.926	0.150	0.068	0.116	0.021
120	22.6	2430	0.067	0.952	0.109	0.098	0.144	0.014
121	29.1	3960	0.102	0.954	0.106	0.130	0.178	0.028
122	19.1	1340	0.050	0.983	0.085	0.069	0.102	0.017

differ by at least 20 percent. As can be seen, the scores for each site are close to unity suggesting that our threshold was set too low. When the threshold was set to a higher value, 2.0, the accuracy scores for Test 4 correspond to results from the other tests. The issue here is not the results--i.e., numbers--themselves; rather, more importantly, the ability to modify a test slightly and alter the outcome completely. This ability gives the model user the flexibility to tailor the test exactly to the issue of concern for his application.

Table 1 shows that the results for Sites 101 and 104 usually account for the highest and second highest scores; the order depends upon the specific test. Other examples of test-specific comparative performance are common in the results presented. For example, observe the results for Tests 6 and 7 at Sites 116 and 120. For Test 6, the accuracy score is lower for Site 116, while the reverse is true for Test 7. The underlying reasons for this type of behavior cannot be explained unless we examine the point-by-point comparisons more closely. This clearly goes beyond the role of the accuracy score, which is to provide a figures of merit (FOM) upon which to base a final evaluation of the model. The point is that our FOMs can mislead a reviewer who fails to understand fully the criteria upon which the FOM is based.

Averaging Time Characteristics

The accuracy score was applied to RAM to obtain a baseline figure-of-merit for various averaging periods. Two accuracy scores--mean square error and location of maximum concentration--were applied to the RAMS/RAPS data bases for averaging times ranging from one hour to one year.

The characteristics of the mean square error are displayed in Figure 3 for the composite of all 13 RAPS sites and Sites 101, 104, and 122, which demonstrates the extremes about the composite. Prediction accuracy improves as averaging time increases in all cases. This observation corresponds to that for other models and other investigators, e.g., Draxler (1979) and Venkatram (1979). Hence, we are more or less confirming an expected result in the case of the mean square error. Figure 4 demonstrates another interesting characteristic of the RAM--improvement in the ability to predict the location of maximum concentration as averaging time increases. Hence, performance improves in two respects as averaging time is increased:

- Improved site-by-site accuracy
- Improved ability to predict the location of maximum concentration.

Application of Mean-Square Error to Two Versions of RAM

After reviewing tests for randomness, it was postulated that the wind interpolating routine was one possible cause. Hence, the wind interpolation scheme was modified and a new RAM/RAPS data base was formed. The difference in the original wind algorithm and modified one are as follows:

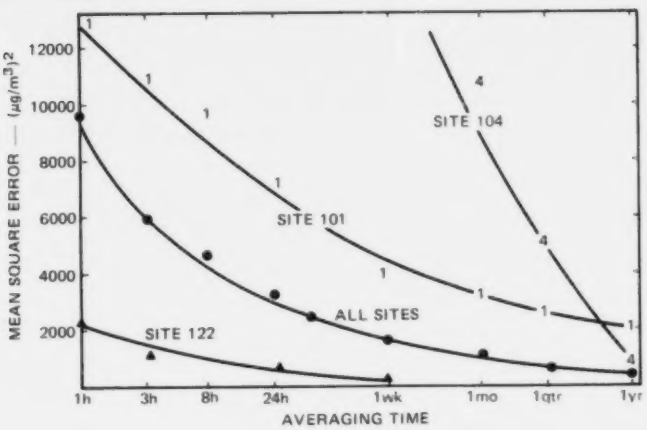


Figure 3. Effects of averaging time on mean square error.

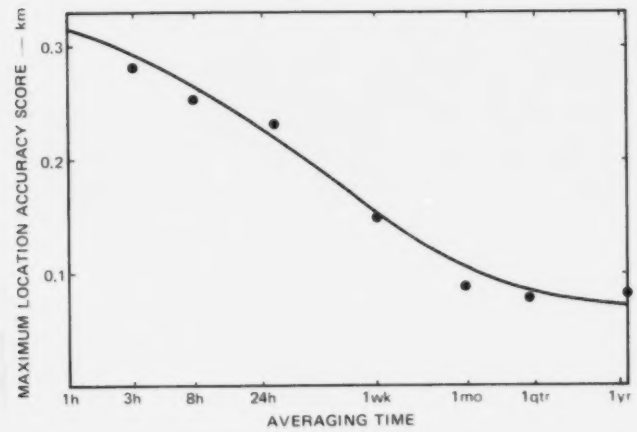


Figure 4. Effects of averaging time on accuracy of predicting location of maximum concentration.

Original	Modified
(1) Interpolates "average" wind speed and direction from all 25 RAPS monitoring stations	(2) Interpolates "average" wind speed and direction from all 25 RAPS monitoring stations
(2) Contains a randomizing wind-direction algorithm, which can change interpolated direction by as much as 10°	(2) Does not contain any further wind-direction adjustments
(3) Computes vector average wind speed	(3) Computes scalar average wind speed

The main difference is the deletion of the randomizing feature, which is normally used with the RAM (Turner and Novak, 1978).

Comparisons of performance for both methods of wind-field interpolations using the mean-squared error FOM were made for both 8- and 24-hour averaging times (see Table 2). No statistically significant differences could be detected in the accuracy scores for the two methods; that is, the accuracy score is not significantly lower for either method because the mean values for the one method are always within the 95 percent confidence intervals of the mean values for the other method. For example, note the confidence intervals

TABLE 2. ACCURACY SCORE (MEAN-SQUARED ERROR) COMPARISON BETWEEN TWO WIND-AVERAGING SCHEMES

Site	8-hour average		24-hour average	
	Original	Modified	Original	Modified
101	9600	9500	6600	6500
103	2300	2400	1200	1200
104	35000	34000	23000	23000
105	2800	2800	1500	1500
106	3700	3600	1900	1900
108	1400	1300	5200	5200
113	1200	1100	590	540
114	1800	1800	800*	850†
115	2000	2500	170	200
116	1200	1200	750	750
120	540	540	250	220
121	110	100	540	530
122	730	720	370	370

*95% confidence interval is 483 to 1110.

†95% confidence interval is 538 to 1170.

beneath Table 2 for the two observations in question; the accuracy score for the modified method, 850, is well within the 95 percent confidence interval for the original, 483 through 1110. Because all comparisons follow this example, we conclude that the modified method is no better than the original method of interpolating winds.

Implications of these results with respect to the RAM are discussed elsewhere. The discussion here is intended to provide the reader with some ideas on how the FOMs might be used to compare models or versions of the same model.

APPLICATION OF DIAGNOSTIC METHODS

Applying all the diagnostic tests at once to the data base produces an overwhelming quantity of results, not all of which are of immediate interest. Our procedure is to apply only a few tests at first to detect bias and randomness as described earlier. We then select additional tests to answer specific questions concerning the model's performance.

Detection of Bias

One very quick way of observing overall bias is to display accuracy score results or observed and predicted concentrations on a geographical map as shown in Figure 1. Note that the agteement is good for some sites (e.g., 106 and 116) and not so good for others (e.g., 101 and 104). Site 101 overpredicts by a factor of two, while Site 104 underpredicts by about 30 percent. This analysis gives us a quick identification of bias for each site. The next step is to examine the frequency distributions for each site. (For a single figure of merit on how well the distributions agree, the chi-square test could also be applied. As discussed later, this chi-square statistic can serve as a basis for comparison on how subsequent model modifications have improved agreement between observed and predicted frequency distributions.) For our purposes here, we will rely on visual displays of the cumulative frequency distribution on logarithmic axes. The logarithmic axes are preferred over linear ones because it enables us to better examine the agreement between observed and predicted high concentrations (extreme values), which occur infrequently (e.g., 0.1 percent of the time).

Logarithm-frequency distributions are displayed* for all sites in Appendix B. Distributions for all sites tend to fall within the extremes of agreement represented by the plots in Figure 5. For Site 101, the observed and predicted concentrations differ almost uniformly throughout the distribution. On the other hand, the crossover for Site 104 indicates the RAM tends to overpredict at the lower concentrations and underpredict at high concentrations. Finally, the data for Site 122 demonstrate good agreement between observed and predicted concentrations.

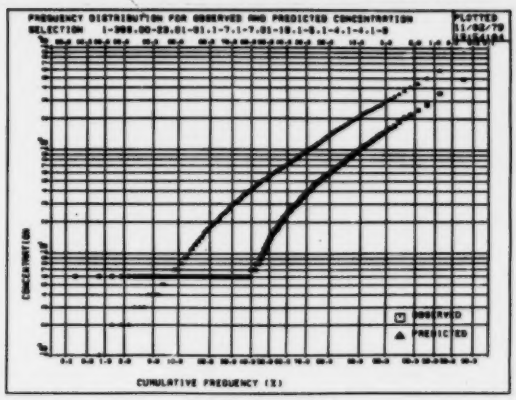
Based on the evidence discussed above, there appears to be no general bias between observed and predicted concentrations. However, there does

---

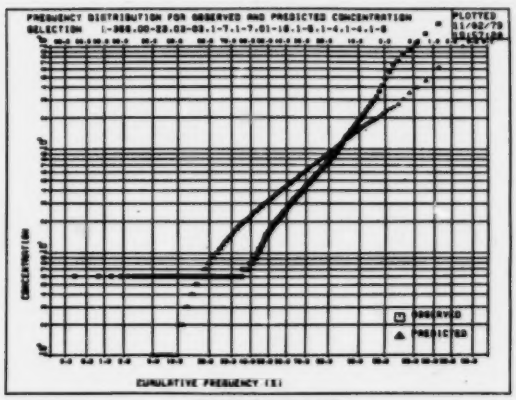
*These displays were produced on a Tektronix 4014 terminal connected to EPA's UNIVAC 1110 facility at Research Triangle Park.

SO_2 FREQUENCY DISTRIBUTIONS FOR 1976 HOURLY AVERAGED OBSERVED AND PREDICTED CONCENTRATIONS ($\mu g/m^3$) FOR SELECTED RAPS SITES

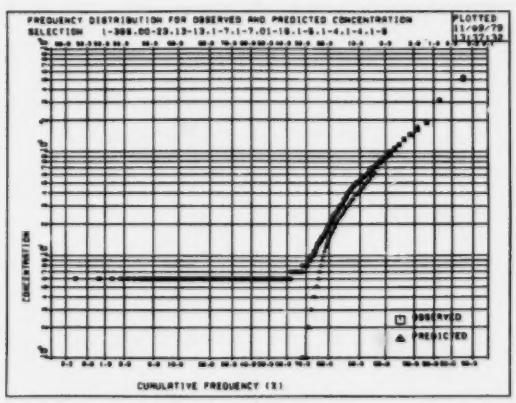
The maximum concentration in the RAPS data base is 6 $\mu g/m^3$



(a) Station 101



(b) Station 104



(c) Station 122

Figure 5. Logarithm frequency distributions for selected sites.

appear to be some local (perhaps spatial) bias, particularly in the vicinity of Site 101. The interstation error correlation test can help diagnose spatial bias. Figure 6 illustrates the results of this test. Figure 6(a) displays the correlation coefficients among the 13 sites; the 95 percent confidence limits are displayed in Figure 6(b).

In reviewing the interstation error correlation coefficients presented in Figure 6, we find that errors at certain sites near Site 101 are moderately correlated to errors at Site 101. (The correlation coefficients between errors at Site 101 and Sites 103, 105, and 106 are 0.51, 0.54, and 0.43, respectively.) We also note that the errors for Site 104 do not correlate well with these other sites. The only other moderate correlation for sites not mentioned exists between Sites 115 and 116 (a correlation coefficient of 0.47). However, examination of the frequency distribution for Sites 115 and 116 does not reveal much bias.

Analysis of the interstation error correlation test in concert with other results leads to the following conclusions:

- Spatial bias exists in the vicinity of Sites 101, 105, 106, and 107. Isolating the cause of errors at one of these sites may have high payoffs at the other sites as well.
- Results for Site 104 do not follow the spatial trend. Some of the causes for rather large errors at this site are not related to causes for errors at other sites (i.e., a highly localized problem).

Detection of Randomness

An effective display for randomness is the scatterplot with regression line and confidence limits displayed with associated correlation and regression coefficients. The hourly case-by-case scatterplots for selected sites are given in Appendix C. Points within the box about the origin were not plotted to save computer costs and to prevent obscuring the point of intercept of the least-squares line for those cases with intercepts close to the origin. Normally, the box would be solid with points, which are not worth the extra computer time to plot. This is perhaps best illustrated with the example of Figure 7. Figure 7(a) gives the 3-hour average case-by-case scatterplot for the composite of all sites excluding the points where both predicted and observed concentrations fall within the 50×70 $\mu g/m^3$ box; Figure 7(b) covers just those cases within the box. In a statistical sense, moderate correlation is said to exist when the correlation coefficient, R, is 0.5 or higher. When considering the 3-hour average for all sites, we say that the comparison is quite random because R is only 0.22 [Figure 7(a)]. Parameters including R, the number of observations (N), and the regression coefficients--slope (M) and intercept (B)--are presented in the lower left-hand corner of the scatterplots.

As it happens, the randomness for the above case is representative of that for hourly averages at any of the sites. The highest correlation coefficient anywhere was 0.24 (Site 104) and the lowest was 0.10 (Site 116). Hence, the observations and predictions are essentially uncorrelated.

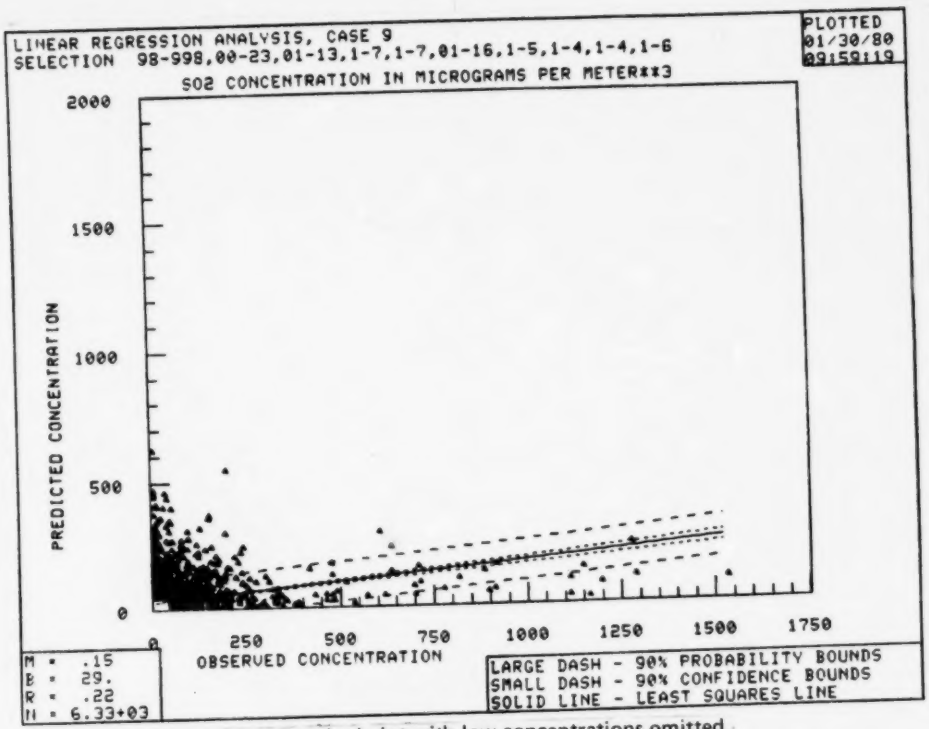
```
INTER-STATION ERROR CORRELATION TESTS.

ENTER DATA ELEMENT ADDRESS OR 11 FOR FILE
ENTER OPTION 1 OR 2
SELECTION 98-998,00-03,01-13,1-7,1-7,01-16,1-6,1-4,1-4,1-6

NO. OF DATA:  586     NO. OF SITES:  13     IOPT:  2     SITE1:  0     SITE2:  0


CORRELATION MATRIX ...

SITE    1    2    3    4    5    6    7    8    9   10   11   12   13
  1  1.00  .61  .23  .54  .43  .33  .28  .18  .21  .12  .08  .20  .11
  2   .61 1.00  .27  .36  .27  .41  .26  .15  .08  .08  .10  .23  .10
  3   .23  .27 1.00  .33  .24  .14  .10  .21  .18  .06  .17  .20  .15
  4   .54  .36  .33 1.00  .44  .33  .23  .15  .16  .05  .12  .21  .12
  5   .43  .27  .24  .44 1.00  .29  .50  .34  .06 -.03  .19  .30  .22
  6   .33  .41  .14  .33  .29 1.00  .25  .20  .07  .07  .04  .16  .27
  7   .28  .26  .10  .23  .50  .25 1.00  .29  .04  .02  .32  .30  .30
  8   .18  .15  .21  .15  .34  .20  .29 1.00  .11  .11  .35  .11  .15
  9   .21  .08  .18  .16  .06  .07  .04  .11 1.00  .47  .11  .01  .03
 10   .12  .08  .06  .05 -.03  .07  .02  .11  .47 1.00  .08 -.02  .08
 11   .08  .10  .17  .12  .19  .04  .32  .35  .11  .08 1.00  .22  .23
 12   .20  .23  .20  .21  .30  .16  .30  .11  .01 -.02  .22 1.00  .28
 13   .11  .10  .15  .12  .22  .27  .30  .15  .03  .08  .23  .28 1.00

95-PERCENT CONFIDENCE LIMITS MATRIX ...
45:>
```

(a) Correlation matrix

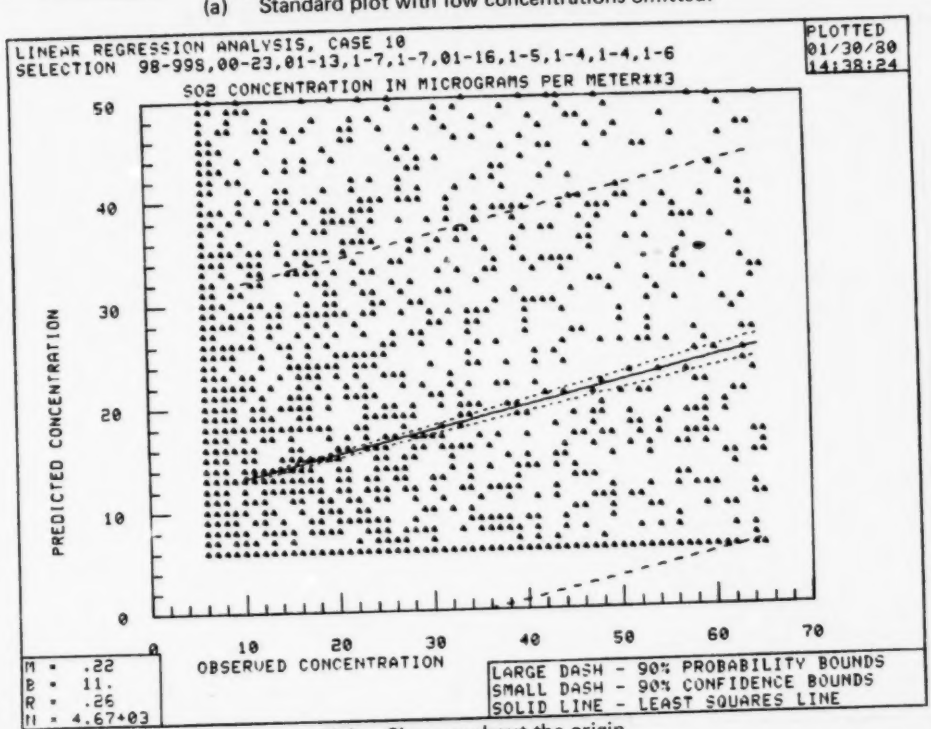
```
95-PERCENT CONFIDENCE LIMITS MATRIX ...

SITE     1     2     3     4     5     6     7     8     9    10    11    12    13
 1 L  1.00   .42   .14   .47   .25   .04   .10   .08   .11   .02  -.08   .11   .01
   U  1.00   .68   .38   .61   .61   .41   .37   .07   .36   .08   .17   .05   .20
 2 L   .42  1.00   .18   .28   .18   .36   .34   .17   .06  -.01  -.01   .01   .15   .01
   U   .68  1.00   .36   .48   .36   .48   .34   .24   .17   .17   .18   .32   .19
 3 L   .14   .18  1.00   .25   .15   .05   .01   .18   .09  -.03   .07   .11   .05
   U   .38   .36  1.00   .48   .33   .23   .19   .30   .27   .15   .26   .29   .24
 4 L   .47   .28   .25  1.00   .36   .25   .14   .07   .07  -.04   .03   .18   .02
   U   .61   .42   .42  1.00   .51   .41   .31   .24   .26   .14   .21   .29   .21
 5 L   .25   .18   .15   .36  1.00   .20   .42   .05  -.04  -.18   .09   .21   .13
   U   .61   .36   .33   .51  1.00   .37   .57   .42   .15   .07   .28   .30   .31
 6 L   .04   .34   .05   .25   .20  1.00   .16   .12  -.02  -.02  -.05   .07   .19
   U   .41   .48   .23   .41   .37  1.00   .33   .29   .16   .16   .13   .25   .36
 7 L   .10   .17   .01   .14   .42   .16  1.00   .01  -.05  -.07   .24   .22   .21
   U   .37   .34   .19   .31   .57   .33  1.00   .38   .13   .12   .40   .39   .38
 8 L   .08   .06   .18   .07   .25   .18   .21  1.00   .02   .02   .27   .02   .06
   U   .27   .24   .38   .24   .42   .29   .38  1.00   .26   .19   .43   .28   .24
 9 L   .11  -.01   .09   .07  -.04  -.02  -.05   .02  1.00   .39   .02  -.09  -.07
   U   .38   .17   .27   .25   .16   .16   .13   .28  1.00   .53   .28   .18   .12
10 L   .02  -.01  -.03  -.04  -.18  -.02  -.07   .02   .39  1.00  -.01  -.11  -.01
   U   .22   .17   .15   .14   .07   .16   .12   .19   .53  1.00   .17   .07   .17
11 L  -.08   .01   .07   .03   .09  -.05   .24   .27   .02  -.01  1.00   .13   .14
   U   .17   .18   .26   .21   .28   .13   .40   .43   .28   .17  1.00   .38   .32
12 L   .11   .15   .11   .12   .21   .07   .22   .02  -.09  -.11   .13  1.00   .19
   U   .29   .32   .29   .29   .39   .25   .38   .28   .18   .07   .38  1.00   .37
13 L   .01   .01   .05   .02   .13   .19   .21   .06  -.07  -.01   .14   .19  1.00
   U   .28   .19   .24   .21   .31   .36   .38   .24   .12   .17   .32   .37  1.00
GBRKPT PRINT8
EOF:87 SCAN:42
01>


CASE D.HOUR3BY6
```

(b) 95-percent confidence limit matrix

Figure 6. Interstation error correlation results. Site numbers on computer listing correspond to RAM station numbers as follows: 1 - 101; 2 - 103, 3 - 104, 4 - 105, 5 - 106, 6 - 108, 7 - 113, 8 - 114, 9 - 115, 10 - 116, 11 - 120, 12 - 121, 13 - 122.



(a) Standard plot with low concentrations omitted.



(b) Close-up about the origin.

Figure 7. Scatterplot for three-hour averaged predicted values versus observed concentrations.

Selection of Specific Problem Areas

In this section, we illustrate how to proceed in our attempt to diagnose specific causes of poor RAM performance. Having determined that there is bias for some sites and not for others and randomness everywhere, we now identify specific problem areas and address one of these to illustrate how additional tests might help us isolate a source of inaccuracy.

Let us consider the worst bias in the network, viz., Site 101, which typically overpredicts by about a factor of two. If we explain the reason for this inaccuracy, we have a good start in improving the overall performance of the model. In fact, because of the results from the interstation error correlation test, we expect substantial improvements at Sites 103, 105, and 106. For Site 104, however, we have observed that separate treatment is probably needed. Hence, in the analysis that follows, we focus separately on Sites 101 and 104.

Diagnosing the Causes of Bias at Site 101

The multiple regression of error residual and discriminant analysis tests were applied to the data base using error residuals at Site 101. Both of these SPSS routines (Nie et al., 1975) considered separately might have led to the same conclusions. These tests are discussed separately in the following paragraphs.

A regression analysis was performed for each of four 90° wind sectors. In this test, the error residual becomes the dependent variable, with the model input parameters as independent variables. Substantial correlations resulted; the highest correlation coefficient (R) was 0.83 for the northwest wind direction sector. This means that about 70 percent* of the residual variance results from the variance in the model input parameters. This result encouraged us to look at even narrower wind sectors. Another regression analysis revealed that the multiple correlation coefficient (R) was about 0.90 for a broad wind direction sector ranging from 270° (W) to 318° (NW). Figure 8 contains the computer-generated† summary table for a wind sector from 279 to 280°. The variable representations are:

RESID is the error (difference in observed and predicted concentrations);

ASC is the Pasquill-Gifford atmospheric stability class;

WSC is the wind speed;

PE is the aggregated point source emissions;

MH is the mixing height;

T is the temperature;

---

*This is given by the square of the multiple correlation coefficient, expressed in percent (i.e., 100 R^2).

†This example and other line printer output to follow were produced at EPA's UNIVAC 1100 facilities in Research Triangle Park.

AE is the aggregated area source emissions;

WDC is the wind direction;

and the column representations are:

MULTIPLE R is the multiple correlation coefficient;

R SQUARE is the square of the above quantity;

RSQ CHANGE is the change in R SQUARE caused by adding the indicated VARIABLE;

SIMPLE R is the bivariate correlation coefficient between RESID and the indicated VARIABLE:

B is the VARIABLE coefficient in the regression equation;

BETA is the normalized version of B.

The corresponding regression equation simplifies to:

$$RESID = 600 - 71 * ASC + 2.6 * WSC - 0.14 * PE + \epsilon$$

where ϵ is a residual term that equalizes the equation.

GLOBAL STATION 1 11/14/80 PAGE 11

FILE RAPSDAT1 (CREATION DATE = 11/14/80)

* MULTIPLE REGRESSION * * * * * * * * * * * * VARIABLE LIST 1
REGRESSION LIST 1

DEPENDENT VARIABLE.. RESID RESIDUAL

SUMMARY TABLE

VARIABLE	MULTIPLE R	R SQUARE	RSQ CHANGE	SIMPLE R	B	BETA
ASC	.77544	.60131	.60131	-.77544	-.7118702+002	-.79820
WSC	.82005	.67248	.07117	.47980	.2604868+001	.50985
PE	.89876	.80776	.13529	-.23272	-.1470220+000	-.88552
MN	.90445	.81802	.01026	.47929	-.2600597-001	-.15435
T	.91010	.82828	.01025	.35479	-.6427955+001	-.45797
WDC	.91035	.82873	.00045	.12775	-.7820908+000	-.02142
AE	.91038	.82880	.00007	.13938	-.2141719-001	-.00916
(CONSTANT)					.6002892+003	

Figure 8. Multiple regression summary table for a wind sector from 279 to 288° – Station 101.

The atmospheric stability class dominated the correlation with a coefficient of -0.78. (In other words, atmospheric stability accounts for almost 60 percent of the variance in the error.) Hence, the model tends to overpredict more with increasing stability over a rather broad northwesterly wind sector. The SCATTERGRAM procedure in the SPSS library was used to graphically illustrate this dependence over the 279° to 288° wind sector; the computer-generated line-printer results appear in Figure 9.

In the discriminant analysis, the errors were separated into three categories: overprediction by 300 $\mu g/m^3$ or more, underprediction by 300 $\mu g/m^3$ or more, and agreement within 300 $\mu g/m^3$. Statistics were computed for model input parameters for each group of error categories. To summarize,

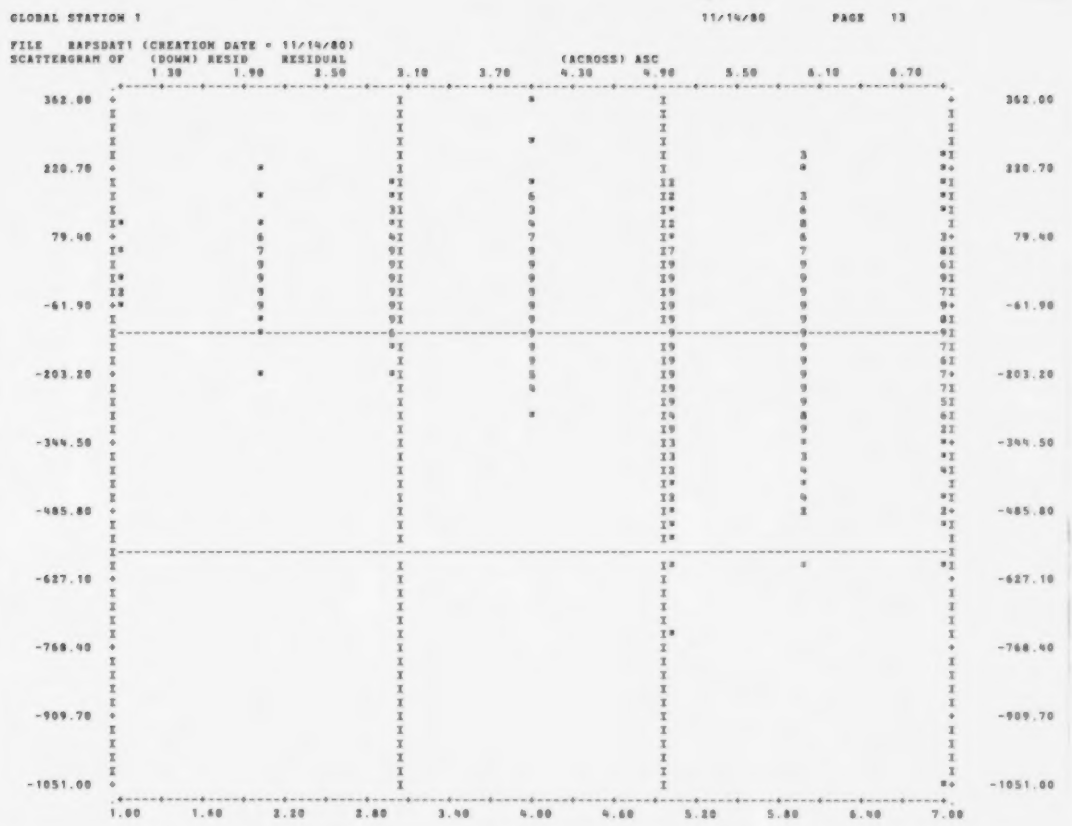


Figure 9. Scattergram of error residual concentrations versus atmospheric stability class (Station 101; wind directions from 279° to 288°).

overpredictions occurred most frequently when winds were westerly, conditions were generally stable, wind speeds were about 1.5 m/s lighter than normal, and point-source emissions were about twice the average for the year. In this case, the discriminant analysis was not as useful as the regression analysis in developing a theory about the causes of overprediction at Site 101. (However, at other sites, the reverse was true in that the discriminant analysis proved to be more useful.)

After analyzing the results, we hypothesized that the area sources immediately to the west of Site 101 were causing overpredictions, particularly during periods of high atmospheric stability. This speculation seemed reasonable because:

- The overprediction occurs over a broad wind sector; a specific point-source influence would occur over a narrower sector.
- At high stabilities and low release heights typical of area sources, the area source "plume" as modeled would not diffuse significantly before touching down at the site.

To test this theory for consistency, we checked the area source inventory to see which grids had high emissions and what their effect on Site 101 might be. As it happens, the area source emissions are highest in the vicinity of Site 101. In fact, the westward adjacent grid emission has the highest in the inventory at 400,000 kg of SO_2 per year. Using techniques outlined by Turner (1970), we estimated that, for a neutral stability, the ratio of concentration to wind speed at Site 101 is 67 $\mu g\text{-}s/m^3$. This concentration, together with contributions from other nearby area sources, could yield a concentration consistent with the annual overprediction of 43 $\mu g/m^3$.

Based on the above analysis, we speculate that lumping a number of small commercial sources into area sources could cause an overprediction. In reality, these small sources would have less impact at Site 101 if they were modeled as individual point sources, rather than as one area source with a large emission rate. The next step, of course, would be to obtain the raw emissions data to see if our hypothesis is correct.

Diagnosing the Causes of Bias at Site 104

As shown in the results from the interstation error correlation test discussed earlier, Site 104 is somewhat of an anomaly in that its residuals do not appear to be related to those from other nearby sites. Hence, rectifying the bias at Site 101 might also rectify some bias at Sites 103, 105, 106 and others, but perhaps not as much at Site 104. Therefore, it seems reasonable to attempt to diagnose the causes of poor performance at Site 104 independently.

The SPSS multiple regression analysis was conducted. However, correlations were much lower than they were for Site 101; in fact, correlation was almost negligible. The tests were applied for four 90° wind sectors (as they were for Site 101). The highest multiple correlation coefficient was 0.43 for the northwest sector; the correlation coefficient was 0.31 or less for the other sectors. Reviewing the results for the northwest sector, the RESID term is negative on the average (showing an average overprediction of about 26 $\mu g/m^3$). For this case, the highest single correlation was between RESID and the atmospheric stability (ASC); this was a negative correlation with a value of -0.26. Because correlations were low, we decided to use a discriminant analysis to evaluate relationships between RESID and the independent variables in another light.

Several discriminant functions were examined. All entailed separation of the residuals into three classes:

- Group 1: Agreement between the observed and predicted concentrations within preset limits (i.e., a symmetric threshold about the perfect agreement where RESID = 0).

- Group 2: Underprediction by more than the preset threshold
- Group 3: Overprediction by more than the preset threshold.

Initially, 300 and 600 $\mu g/m^3$ thresholds were tried. Partial computer results for each threshold are given in Figures 10 and 11. The former shows that for the 300 $\mu g/m^3$ threshold, 2251 cases are in agreement, 57 result in underprediction, and 34 result in overprediction; the latter shows that, for 600 $\mu g/m^3$, 2304 cases are in agreement, 30 result in underprediction, and 8 in overprediction.

Examining the tables of parameter means and standard deviations leads to the following general conclusions:

- Underpredictions occur most frequently during slightly stable conditions, low wind speeds (given in tens of m/s), low mixing heights, very low temperature (averaging 0.1333°C), and high point source emissions.
- Overpredictions occur most frequently (in comparison to underpredictions) during more stable, lower wind speeds, and lower mixing heights. Also, estimated point source emissions are much lower than average, and mean temperatures are higher than normal.
- Wind directions do not favor over- or underpredictions.
- The amount of underpredictions is considerably higher than the amount of overprediction (note the RESID values in the MEANS tables in Figures 10 and 11).
- Because the standard deviations (see tables in Figures 10 and 11) are fairly high compared to the means, there are many exceptions to the general conclusions listed above.

The last point above is perhaps best demonstrated by the prediction accuracy of the discriminant function. This result, shown in Figure 12, shows how well the function, using the model input parameters, is able to categorize the cases into the three groups. According to Figure 12, the discriminant function correctly classifies only 59.91 percent of the time. The implication is that there is no well-defined set of meteorological and emissions data that uniquely can be analyzed to identify (or predict) cases of over- or underprediction.

In the above analyses, the average amount of underprediction (about 1058) stands out. (This corresponds to the RESID mean of the 30 cases that were in GROUP2 as shown on Figure 11 for the 600 $\mu g/m^3$ threshold.) Because the number of cases was not large, it was a fairly simple matter to examine the data base on a case-by-case basis. This process revealed that most of the large underpredictions occurred during December favoring no particular time of day nor wind direction and occurring quite often over extended periods of time (e.g., several hours). Figure 13 illustrates one particular extended period.

The analysis was carried further using the CROSSTABS procedure (an option available in the DISCRIMINANT program). This procedure enables one to review

GLOBAL STATION 4 11/14/80 PAGE 2

FILE DBAT10% (CREATION DATE = 11/14/80)

- D I S C R I M I N A N T A N A L Y S I S -

ON GROUPS DEFINED BY ABSDIF

2342 (UNWEIGHTED) CASES WERE PROCESSED.
0 OF THESE WERE EXCLUDED FROM THE ANALYSIS.
2342 (UNWEIGHTED) CASES WILL BE USED IN THE ANALYSIS.

NUMBER OF CASES BY GROUP

ABSDIF	NUMBER OF CASES UNWEIGHTED	WEIGHTED	LABEL
1	2251	2251.0	
2	57	57.0	
3	34	34.0	
TOTAL	2342	2342.0	

GROUP MEANS

ABSDIF	DAY	TM	ASC	WSC	WDC	T	AZ	PZ
1	193.17903	11.65837	4.43447	38.35495	212.84007	10.53932	270.56153	1167.05598
2	337.28070	12.82456	4.78175	30.67719	222.31053	2.66667	250.73684	1760.68421
3	312.70588	8.67647	5.85294	18.14706	213.73529	8.47059	265.47059	1253.11765
TOTAL	196.96968	11.64347	4.46157	37.87959	213.08113	10.31768	270.00513	1192.75330

ABSDIF	MM	RESID	OC	PC
1	930.51755	-18.74234	32.20702	50.94936
2	553.49123	750.91228	865.56140	114.64912
3	395.08824	-487.44118	59.97059	547.41176
TOTAL	913.56832	-6.81469	52.89240	59.70709

GLOBAL STATION 4 11/14/80 PAGE 3

GROUP STANDARD DEVIATIONS

ABSDIF	DAY	TM	ASC	WSC	WDC	T	AZ	PZ
1	109.73418	6.97731	1.23091	19.25713	105.45192	9.98819	92.85529	772.32843
2	31.48625	6.89234	1.17966	16.96748	81.55996	6.78584	39.79928	630.98036
3	106.20949	7.66227	1.10460	8.41403	55.40939	11.65732	39.23375	1016.70802
TOTAL	110.69329	6.99403	1.23975	19.27104	104.36637	10.01357	92.82992	779.31166

ABSDIF	MM	RESID	OC	PC
1	632.39696	68.27881	55.62580	59.26868
2	313.76379	437.85550	449.49496	100.73436
3	454.78489	172.48631	73.04772	198.03583
TOTAL	630.06977	164.18847	156.12525	88.16805

Figure 10. Discriminant group summaries for agreement, underprediction, and overprediction using 300 $\mu g/m^3$ threshold.

Group 1 (agreement): $|OC - PC| \leq 300\ \mu g/m^3$
Group 2 (underprediction): $OC - PC > 300\ \mu g/m^3$
Group 3 (overprediction): $PC - OC > 300\ \mu g/m^3$

GLOBAL STATION 4 11/14/80 PAGE 2

FILE DDAT104A (CREATION DATE = 11/14/80)

- DISCRIMINANT ANALYSIS -

ON GROUPS DEFINED BY ABSDIF

2342 (UNWEIGHTED) CASES WERE PROCESSED.
0 OF THESE WERE EXCLUDED FROM THE ANALYSIS.
2342 (UNWEIGHTED) CASES WILL BE USED IN THE ANALYSIS.

NUMBER OF CASES BY GROUP

ABSDIF	NUMBER OF CASES UNWEIGHTED	WEIGHTED	LABEL
1	2304	2304.0	
2	30	30.0	
3	8	8.0	
TOTAL	2342	2342.0	

GROUP MEANS

ABSDIF	DAT	TM	ASC	WSC	WDC	T	AE	PE
1	194.95964	11.62457	4.44835	38.09158	213.05729	10.43663	270.21484	1174.40972
2	341.90000	13.56667	5.00000	28.03333	214.93333	.13333	258.36667	1972.60000
3	232.37500	9.87500	6.25000	13.75000	213.00000	14.25000	253.25000	623.75000
TOTAL	196.96968	11.64347	4.46157	37.87959	213.08113	10.31768	270.00512	1182.75320

ABSDIF	MN	RESID	OC	PC
1	919.20052	-18.10373	38.04557	56.14931
2	620.83333	1057.66667	1184.36667	126.70000
3	389.25000	-747.37500	85.75000	833.12500
TOTAL	913.56832	-6.81469	52.89240	59.70709

GLOBAL STATION 4 11/14/80 PAGE 3

GROUP STANDARD DEVIATIONS

ABSDIF	DAT	TM	ASC	WSC	WDC	T	AE	PE
1	110.25037	6.98470	1.23487	19.25034	104.78005	10.00744	42.78517	777.66206
2	16.00937	7.08414	1.23176	16.70739	86.00078	4.78311	46.85670	470.55855
3	69.76888	8.99901	1.03510	5.11999	3.02372	4.80327	34.74293	413.35656
TOTAL	110.69329	6.99403	1.23975	19.27104	104.36637	10.01357	42.82992	779.31166

ABSDIF	MN	RESID	OC	PC
1	631.99556	92.56047	77.05040	74.60328
2	340.90954	398.82511	389.62267	106.82088
3	426.71695	113.42453	95.49682	119.64285
TOTAL	630.06977	164.18847	156.12525	88.16805

POOLED WITHIN-GROUPS COVARIANCE MATRIX WITH 2339 DEGREES OF FREEDOM

Figure 11. Discriminant group summaries for agreement, underprediction, and overprediction using 600 $\mu g/m^3$ threshold.

Group 1 (agreement): $|OC - PC| \leq 600\ \mu g/m^3$
Group 2 (underprediction): $OC - PC > 600\ \mu g/m^3$
Group 3 (overprediction): $PC - OC > 600\ \mu g/m^3$

GLOBAL STATION 4 11/14/80 PAGE 16

FILE BDAT104A (CREATION DATE = 11/14/80)

- D I S C R I M I N A N T A N A L Y S I S -

ON GROUPS DEFINED BY ABSDIF

0

ANALYSIS NUMBER.. 1

NUMBER OF CANONICAL DISCRIMINANT FUNCTIONS.. 2

LIST OF THE 4 VARIABLES USED..

VARIABLE	LABEL
ASC	
WSC	
T	
MN	

CLASSIFICATION RESULTS -

ACTUAL GROUP	NO. OF CASES	PREDICTED GROUP MEMBERSHIP 1	2	3
GROUP 1	2304	1374	479	451
		59.6%	20.8%	19.6%
GROUP 2	30	7	22	1
		23.3%	73.3%	3.3%
GROUP 3	8	1	0	7
		12.5%	.0%	87.5%

PERCENT OF 'GROUPED' CASES CORRECTLY CLASSIFIED: 59.91%

CLASSIFICATION PROCESSING SUMMARY

2342 CASES WERE PROCESSED.
0 CASES HAD AT LEAST ONE MISSING DISCRIMINATING VARIABLE.
2342 CASES WERE USED FOR PRINTED OUTPUT.

Figure 12. Prediction accuracy of discriminant function – 600 $\mu g/m^3$ threshold.

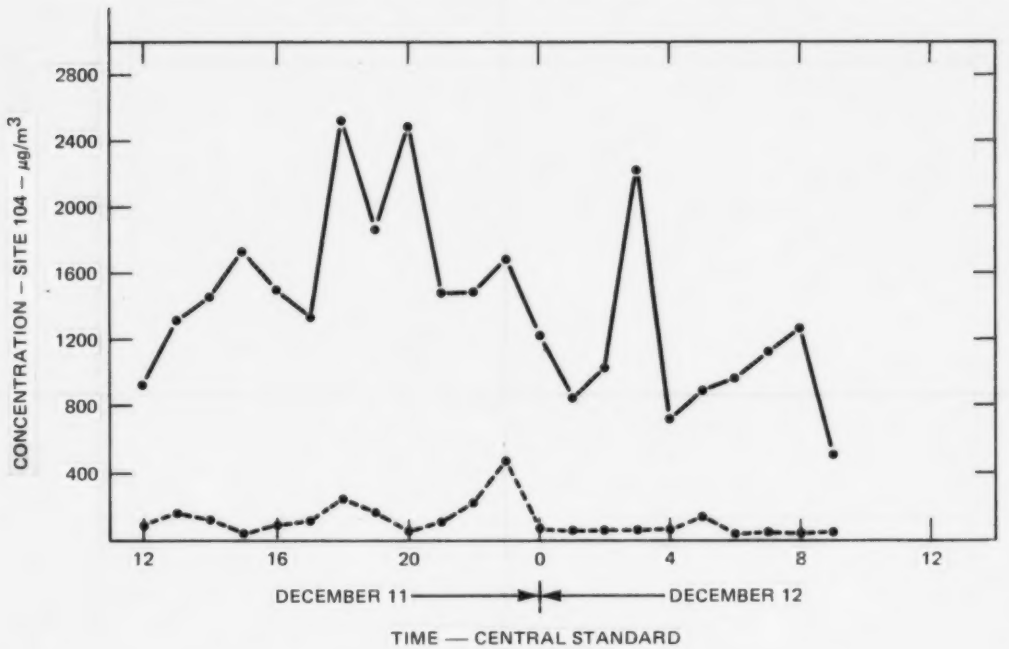


Figure 13. Time sequence of high observed concentrations (solid line) at site 104 during December 1976. (Predicted concentrations are shown by the dashed line)

the amount of misprediction as a function of one or two dependent variables (e.g., wind direction). In our case, we are interested in determining whether there is a predominant wind direction that most often leads to excessive underprediction. Figure 14 illustrates a CROSSTABS output for residual categories as a function of wind speed class and wind direction class. In this case, the residuals were separated into nine 200-µg/m^3 ranges starting at underpredictions of 1000 µg/m^3 (and above) as the leftmost column to overpredictions of 600 to 800 µg/m^3 as the rightmost column. Wind direction was separated into 36 10° segments. The resulting output for a wind speed class of 2 m/s and above is contained on five pages, each with up to eight segments (Figure 14 is the fourth such page). Wind direction categories (WDCLS) are denoted by the numbers to the left of the leftmost residual class (RESCLS); the number multiplied by ten represents the upper limit of the 10° category (note the 290°-300° designator in the lower left-hand corner).

The results show that for the 13 cases involving underpredictions of 1000 µg/m^3 and greater, six of them are represented in the wind sectors given in Figure 14 (i.e., three in the 240° to 250° sector, two in the 280° to 290° sector, and one in the 270° to 280° sector). Table 3 illustrates a complete wind direction breakdown (20° sectors) for the year and separately for December when most large underpredictions occurred.

CROSSTABS 104 11/14/80 PAGE 10

FILE BDAT104A (CREATION DATE = 11/14/80)

* CROSSTABULATION OF *
WDCLS BY RESCLS
CONTROLLING FOR..
WSCLS VALUE.. 2 GE 1MPS
* * * * * * * * * * * * * * * * * * * \

TABLE 3. SITE 104 UNDERPREDICTION OF 600 $\mu g/m^3$ AND GREATER BY WIND DIRECTION (Number of occurrences in SPSS data base)

Wind direction (degrees)	December 1976	Total 1976
0-20	0	0
20-40	1	1
40-60	0	0
60-80	0	0
80-100	0	0
100-120	0	0
120-140	1	1
140-160	2	2
160-180	0	0
180-200	2	2
200-220	0	1
220-240	3	4
240-260	4	4
260-280	4	4
280-300	0	0
300-320	0	0
320-340	0	0
340-360	3	3
Total	20	22

Irwin (1980) carried the CROSSTABS analysis further examining seasonal difference and residuals of 200 $\mu g/m^3$ (both underpredictions and overpredictions). The results of Irwin's analysis are summarized in Table 4 and illustrated more thoroughly in Appendix D. His results are consistent with the findings discussed previously.

The results presented in this section indicate that there are at least two factors influencing the RAM performance at Site 104. Referring to Figure 5 (page 23), it can be observed that for the lower concentrations (which occur most frequently), the model overpredicts and for the high concentrations (which occur infrequently), the model severely underpredicts. The two extremes of the frequency plot could be a manifestation of the following two controlling factors:

- The overprediction of lower concentrations bears many similarities to the overall performance at Site 101. For Site 101, it was postulated that overpredictions occurred because the emission inventory represented too many small individual emission sources as one large area source. Furthermore, the interstation error correlation test demonstrated moderate correlations between Site 101 and other sites in the vicinity. We believe that the lower-concentrations performance at Site 104 would improve if our remedy for Site 101 proved effective.

TABLE 4. IRWIN'S PRELIMINARY CONCLUSIONS FOR SITE 104 (1976)

a. Conditions Favorable for Underpredictions Greater than 200 $\mu g/m^3$

Days 1-335	Days 336-366
1. Favor nighttime	1. No time preference
2. Generally, within any wind sector over and under predictions offset one another in frequency of occurrence	2. Prefer a. 120° - 160° b. 180° - 280°
3. Favor Pasquill-Gifford Class 7 (G)	3. Favor Pasquill-Gifford Class 5 (E)
4. Favor mixing heights 200-800 m	4. Favor mixing heights 200-800 m
5. Favor wind speeds 0-2 m/s	5. Favor wind speeds 1-3 m/s
	6. Largest errors and most frequent occurrence seems to be late November to mid-December

b. Conditions Favorable for Overpredictions Greater than 200 $\mu g/m^3$

Days 1-335	Days 336-366
1. Favor nighttime	1. Favor nighttime
2. Definitely favor wind sector 200-220°	2. Slightly favor wind sector 280-320°
3. Definitely favor Pasquill-Gifford Class 7 (G)	3. Favor Pasquill-Gifford Class 5 (E)
4. Favor mixing heights 200-400 m	4. Favor mixing heights 200-400 m
5. Definitely favor wind speeds $\leq$2 m/s	5. Definitely favor wind speeds 2-4 m/s

- The large underpredictions do not appear to be correlated among sites and hence are an anomaly at Site 104. We believe that the large observed concentrations result from an uncategorized or poorly quantified emission source located in the vicinity of the site (most likely in the south to west sector). As seen in Figure 15, several substantial point sources are nearby. Because the underpredictions occurred over a brief period (mostly in December), we speculate that the error could have resulted from poorly quantified (or neglected) emissions from one of these nearby sources.

We note that both sources of error described above are not inaccuracies in the RAM formulation; rather, they are manifestations of errors (or approximations) in the emission input data.

Comparison of Results from Other Sites with Results from Sites 101 and 104

The multiple regression analysis was conducted for Sites 103, 106, 116, and 122. The rationale was to examine results at locations representing different distances from the center of the city (Site 101 is generally thought of as being the center of the study area). Table 5 demonstrates the results showing, first of all, that Site 101 has the highest multiple correlation (as discussed previously) and demonstrates the low correlations for Site 104. There are moderate correlations for other sites. For instance, Site 103 shows correlations of 0.56 and 0.57 for the northeast and southwest sectors respectively; Site 106 shows 0.60 and 0.61 for the northeast and northwest sectors; and other Sites 116 and 122 have correlations of 0.34 or lower.

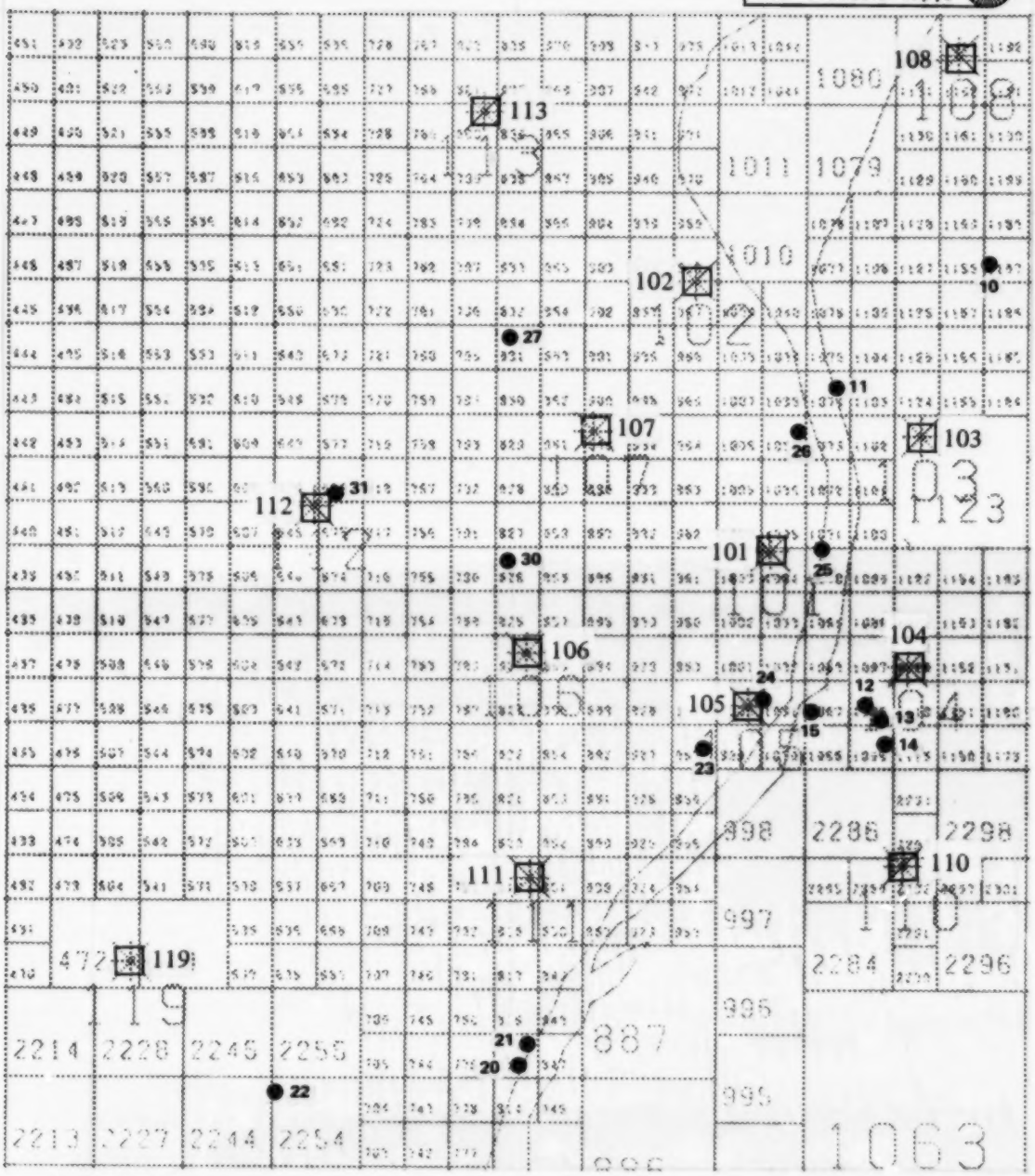
As demonstrated in the interstation error correlation test, the residuals are related among sites. Therefore, rectifying the problem at Site 101 is expected to also improve performance at other sites, particularly Site 103.

In examining the individual correlations, we note that moderate correlations frequently exist between residuals and ASC, PE, and T. The ASC negative correlation has been discussed in relation to the Site 101 and Site 104 performance. (Since the bivariate correlation between PE and T is -0.95, we expect their correlations with RESID to be comparable in magnitude. In future discussions we focus on PE rather than T for the sake of developing a theory on causes.) It is theorized that the high negative correlation with the point source emissions (PE) is related to the problems in estimating wind trajectories (which was theorized to be the primary cause of randomness). If the modeled wind trajectory is substantially in error more often than not, point source emissions could be modeled upwind too often, increasing the predicted concentration in relation to that actually observed (making the residual more negative). We feel this is one primary factor in the degree of randomness exhibited by the RAM.

FUTURE ANALYSIS

Further diagnostic tests are not warranted until further analysis of the extensive RAPS data base is undertaken with the objective of verifying the theories already presented herein. If our theories are correct, then the RAM (or the model input data base) should be modified. These modifications would

Reproduced from best available copy.



POINT SOURCE KEY:

10	GRANITE CITY STEEL	20	N.L. INDUSTRIES	26	PVD INTERNATIONAL
11	UNION ELECTRIC - VENICE	21	GREAT LAKES CARBON	27	GENERAL MOTORS
12	AMAX ZINC	22	ALPHA CEMENT	30	WASHINGTON UNIVERSITY - EUCLID
13	EDWIN COOPER	23	ANHEUSER BUSCH		
14	MONSANTO CO.	24	MONSANTO CO., SECOND ST.	31	WASHINGTON UNIVERSITY CAMPUS
15	UNION ELECTRIC - CAHOKIA	25	UNION ELECTRIC - ASHLEY		

Figure 15. Point source locations (●) in relation to RAPS monitoring stations with a background of area source grids.

TABLE 5. SUMMARY OF MULTIPLE REGRESSION STATISTICS FOR SELECTED SITES

Site	Wind sector	Mean RESID	No. of cases	Multiple R	Simple R		
					Highest	2nd Highest	3rd Highest
101	NE	---	396	0.67	T(0.53)	PE(-0.52)	ASC(-0.32)
	SE	---	468	0.36	T(0.30)	PE(-0.25)	ASC(-0.26)
	SW	---	484	0.65	PE(-0.50)	T(0.50)	ASC(-0.34)
	NW	-86	994	0.83	ASC(-0.57)	PE(-0.48)	T(0.51)
103	NE	-19	396	0.56	WDC(0.41)	WSC(0.22)	ASC(-0.18)
	SE	1	468	0.36	ASC(-0.30)	AE(0.20)	T(0.19)
	SW	-38	484	0.57	ASC(-0.51)	MH(0.36)	WSC(0.34)
	NW	-13	994	0.34	WSC(0.25)	ASC(-0.23)	MH(0.22)
104	NE	-14	396	0.22	ASC(-0.13)	---	---
	SE	12	468	0.31	PE(0.26)	---	---
	SW	3	484	0.29	ASC(-0.18)	---	---
	NW	-16	994	0.37	ASC(-0.20)	---	---
106	NE	-34	396	0.60	T(0.47)	PE(-0.46)	ASC(-0.33)
	SE	5	468	0.33	T(0.27)	PE(-0.23)	ASC(-0.20)
	SW	4	484	0.18	AE(-0.10)	---	---
	NW	-14	994	0.61	WDC(-0.48)	ASC(-0.35)	WSC(0.28)
116	NE	7	396	0.20	ASC(-0.14)	---	---
	SE	3	468	0.34	ASC(-0.28)	---	---
	SW	-8	484	0.34	ASC(-0.23)	---	---
	NW	-8	994	0.23	ASC(-0.21)	---	---
122	NE	10	396	0.16	AE(0.11)	---	---
	SE	13	468	0.24	ASC(-0.20)	---	---
	SW	-4	484	0.26	ASC(-0.20)	---	---
	NW	7	994	0.23	WSC(-0.11)	---	---

consist of improving the manner by which large area sources are treated. Also, the suspected emission estimate errors near Site 104 should be verified and corrected. It is hoped that this would result in a significant improvement. The new version of RAM could then be subjected to further tests. The correction of existing deficiencies would make other problems (with RAM) less obscure and new theories could be developed. Hence, we are in the midst of an iterative process, which could result in stepwise improvements in performance.

In conducting further analysis of the RAPS data base, it is desirable (if not essential) to work with actual measurements in addition to model input parameters. Almost all of the RAM input parameters are modeled themselves (e.g., the wind field is interpolated from a number of wind station measurements; emissions rely on estimated fuel consumptions and emission factors; plume rise relies on unverified parameterizations). There is no assurance that the model input parameters are representative of the actual meteorological and emission processes. Therefore, observed concentrations, assuming the worst case, may not be correlated at all with the input parameters. In analyzing the RAPS data base, one objective should be to determine what meteorological and emissions measurements (or estimates) correlate well with observed concentrations. Then we must determine if those correlated measurements are reasonably integrated into the quantification of the model input parameters.

SECTION 6

ASSESSMENT OF ALTERNATIVE STATISTICS

RELATIVE MERIT AMONG TESTS

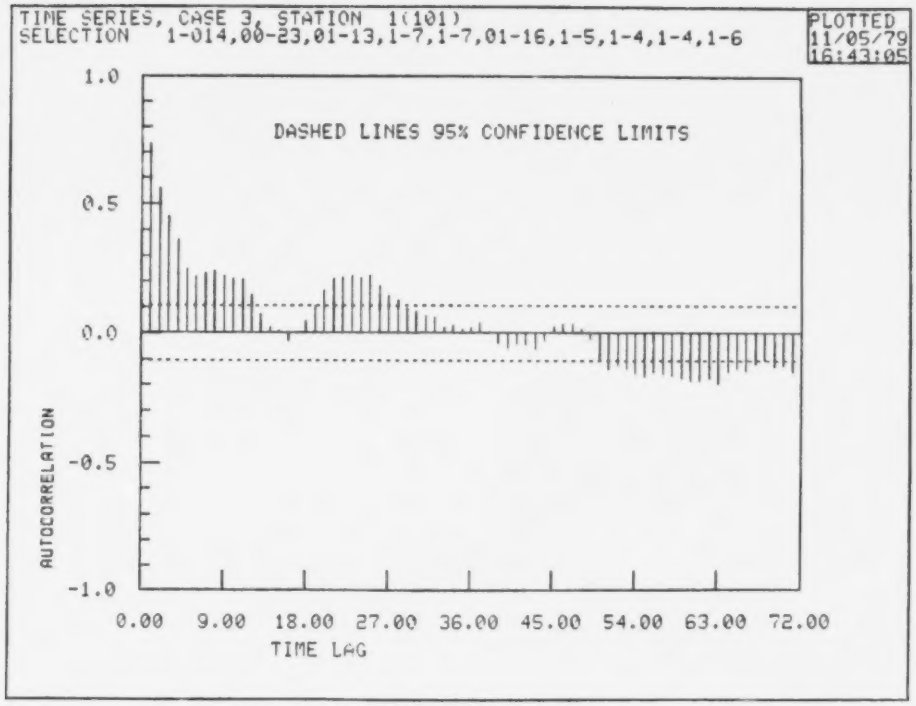
The type of application determines which of our accuracy scores might be most useful. In this case, the performance criteria may vary, depending upon the requirements of the person applying or evaluating the model. This topic has been discussed in our previous work. We note that all of the accuracy scores rely on comparisons of observations at the same place and time. Later we discuss some alternatives to the final evaluation statistics presented herein.

In applying the diagnostic methods, we found that there were many redundancies; that is, tests that produced similar results. For instance, the multiple regression and discriminant analyses could lead to the same conclusions about model performance. However, the discriminant analysis was easier to interpret for a case (for Site 104) entailing a small number of large residuals (mispredictions) while the regression analysis was more useful for a case (Site 101) where residuals were small but strongly biased toward general overprediction. This example suggests that it may be difficult to assess the relative utility or merit of the diagnostic tests. In other words, the particular application determines which test is most useful.

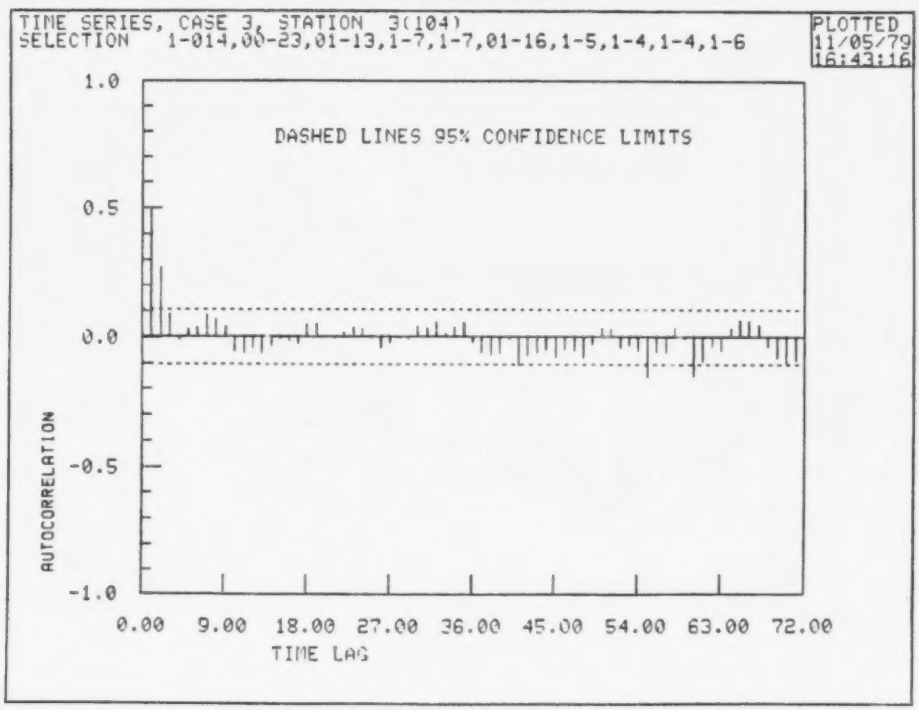
Two statistical tests (or portions thereof) were not used in our analysis, although they were applied: the chi-square statistic and associated displays and the time series of error residuals. The chi-square statistic is considered in the next subsection as a potentially useful final evaluation statistic. The time-series results were not easy to relate to causes of misprediction.

Two examples of time-series results, one each for Sites 101 and 104, are presented in Figure 16. For both sites, autocorrelations are highest for short time lags (e.g., one to two hours). For Site 101, the correlations start out high--over 0.7 for a lag of one hour--and decrease to about one-half that after four hours. After 12 hours, the correlation drops to 0.1, but peaks again after 24 hours. This means there is a slight correlation between the errors occurring at the same time on succeeding days. The correlation drops off and increases slightly as the time lag reaches 50 hours. For Site 104, the correlation is generally 0.1 or below after two hours. Summarizing the results in more meaningful terms:

- The errors at Site 101 are related to the same causes from hour to hour and to a lesser extent from day to day (e.g., lags of 24 hours). For lags of three hours and longer, the error correlation is quite low.



(a) Station 101.



(b) Station 104.

Figure 16. Autocorrelation of error residual versus time lags (in hours).

- The errors at Site 104 are related from hour to hour but essentially random for longer time lags. This suggests a random cause and is consistent with our theory that the causes are uncatalogued emissions, occurring on a nonuniform schedule.

The above findings support our previous findings, but by themselves are not sufficient to develop theories about causes of mispredictions. However, the time-series tests could prove useful in other model evaluation activities.

ALTERNATIVE FINAL EVALUATION STATISTICS

Other Research

Final evaluation statistics have been the subject of ongoing investigations over the past few years. Since our original survey in early 1978, others have reported final evaluation statistics worthy of consideration. Work by Hayes (1979), Bowne (1980), and Martinez et al. (1980) are examples of efforts that strive to define comprehensive model evaluation criteria. In Hayes' report, five desired model attributes were defined as:

- Spatial alignment of concentration patterns
- Temporal correlation
- Accuracy of peak prediction
- Lack of gross error
- Absence of systematic bias.

Bowne extends Hayes' works by recommending a formal model validation protocol. Current work by Martinez et al. (1980) recommends attributes similar to those described by Hayes while combining separate tests for the attributes into a single figure of merit (FOM) using a scale of one to ten.

Our earlier report (Javitz and Ruff, 1980) considered using the chi-square test in addition to our accuracy scores as a final evaluation statistic. The chi-square test, discussed in the next section, is valid only in evaluating the frequency distributions between observed and predicted concentrations, and coupled with information on extreme values, should be quite useful for regulatory purposes.

Using Attributes of the Frequency Distributions as a Criteria

Javitz and Ruff (1980) suggested the use of a chi-square test as a final evaluation statistic for those applications when the comparison between frequency distributions for observed and predicted concentrations is of primary concern. This type of comparison addresses the question of how well a model predicts a number of predefined concentration ranges for various averaging times (e.g., 3-hour, 24-hour) over an extended period of time (e.g., a year); it does not tell us how well the model will perform on a real-time basis, as required for emergency episodes. Results from the chi-square tests include a chi-square statistic and displays of observed and predicted frequency distributions, error residuals, and percentage error residuals. However, this

information will still fall short of answering certain critical issues concerning performance at extreme values.

As it stands, the chi-square test divides pollution classes into rather broad ranges, with all values above a certain concentration falling into the last range. This tells us nothing about the actual magnitude of the misprediction at the extreme values. Therefore, we recommend tabulating separately several highest predicted and observed concentrations. This information, coupled with results from the chi-square tests, provides information more suitable for evaluation of the frequency distributions and the associated extreme values.

Chi-Square Goodness-of-Fit Test--

The chi-square goodness-of-fit test can be used to assess whether the distribution of predicted concentrations deviates significantly from the distribution of observed concentrations, and to depict such departure graphically. Differences between the observed and predicted concentration distributions, such as different variances or shapes, may be identified from appropriate data displays.

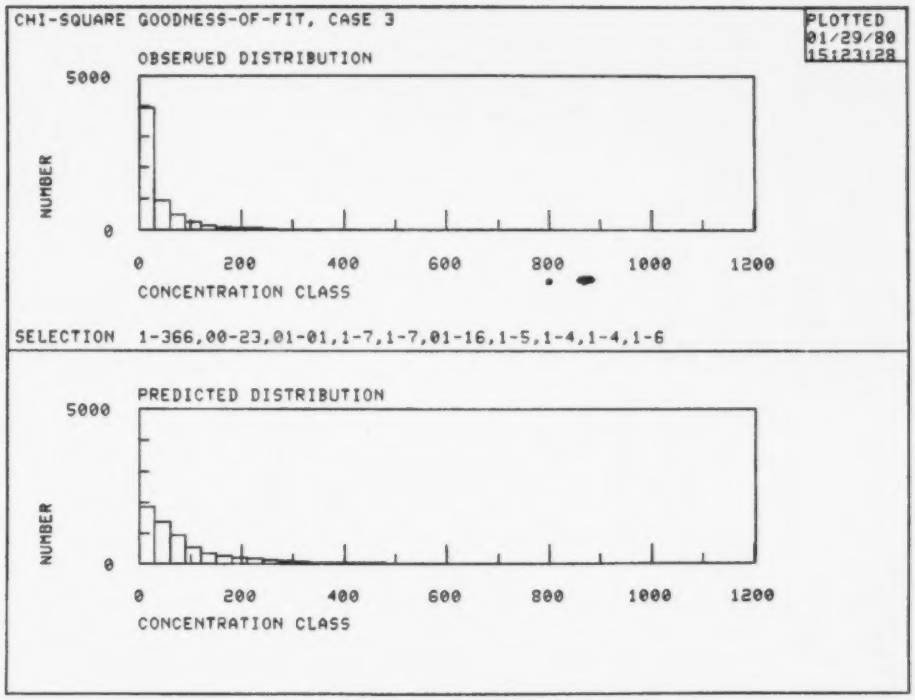
The observed concentrations, $OC_1,\ldots,OC_n$, are used to define cell boundaries for air-pollution concentration with an approximately equal number of observed concentrations in each cell. If possible, each cell should have approximately ten observations. By E_j, denote the number of observed concentrations in the jth cell, where there are J cells, and $1 \leq j \leq J$. By P_j, denote the number of predicted concentrations of the jth cell. If the model is accurate, the E_j and the P_j should be approximately equal.

Under the assumption that the E_j are the expected number of predicted concentrations in the jth cell, the statistic

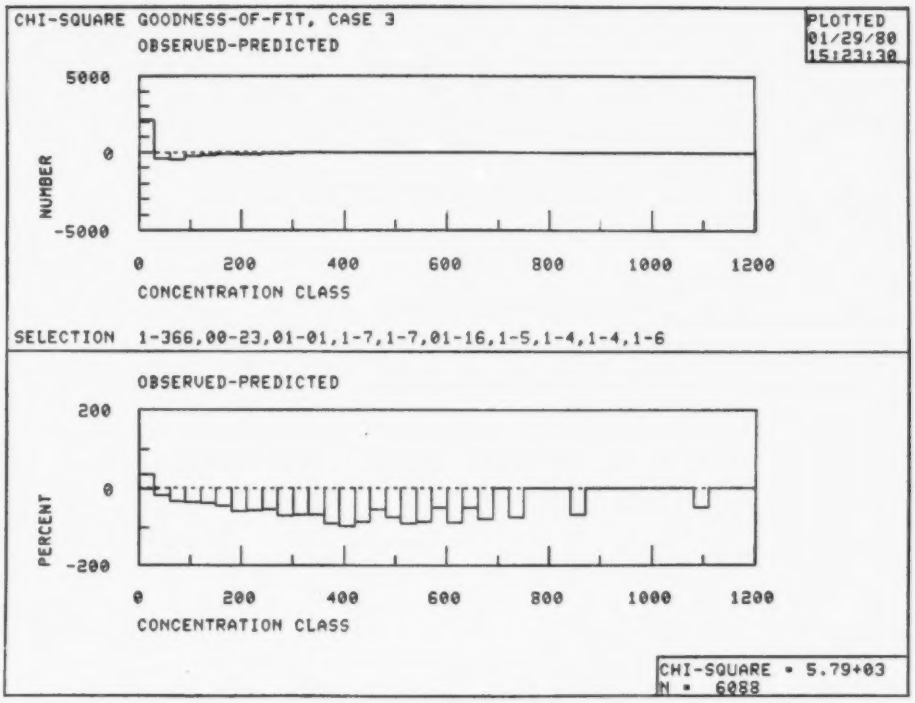
$$\chi^2 = \sum_{j=1}^{J} \frac{(P_j - E_j)^2}{E_j}$$

is distributed as a chi-square random variable with J - 1 degrees of freedom.

Larger values of the chi-square statistic than might reasonably be anticipated from a chi-square random variable are indicative of poor model fit. The significance of the chi-square statistic can be obtained from a chi-square distribution table. It is usual practice to report the value of the chi-square statistic (χ^2) as well as its significance (obtained from appropriate tables). In our case, we chose to augment the statistic with displays as shown in Figure 17. The example, for Site 101, consists of four plots. The first two, Figure 17(a) are simply the frequency distribution for the observed and predicted concentrations, respectively. The second two show the distribution of the residuals and the percentage residual. The chi-square statistic is presented in the lower right-hand corner. Referring to Figure 17(a), we see the limitations of a linear scale in illustrating extreme (high concentration) values. (For this reason, we feel that log-frequency plots are more useful.) The percentage residual plot [Figure 17(b)] does indicate the presence of



(a) Observed and predicted concentrations



(b) Absolute and percent residuals

Figure 17. Frequency distribution by class on a linear scale for Station 101.

large mispredictions. In the lower panel of Figure 17(b), we observe that concentrations in the 1100-$\mu g/m^3$ category were overpredicted about 50 percent of the time.

Table 6 presents values for the chi-square statistic for selected sites. If we consider this statistic as our final evaluation criteria, Site 101 appears to be the worst performer and Site 122 the best. Other than as a basis for comparison among sites, the χ^2 statistic can be used to test the hypothesis that the predicted and observed distributions are the same. Chi-square percentile tables indicated that, for the hypothesis to be true at a 0.01 significance level, the value of χ^2 must be about 60 or less (for 39 degrees of freedom). We can conclude that, for those sites listed, the observed and predicted concentration distributions are probably different. Site 122 comes the closest but still does not pass the test. Comparisons of distributions among sites are probably best done visually. As demonstrated in Figure 5 and elsewhere, Site 104 is the worst performer with respect to predicting high concentrations; however, the value for χ^2 is much higher for Sites 101 and 103. Hence the χ^2 value would be misleading if we are most concerned about performance at high concentrations. Hence, it is desirable to supplement the histograms and the chi-square statistic with tables of observed and predicted high concentrations, as discussed below.

TABLE 6. VALUE OF CHI-SQUARE AS A FUNCTION OF SITE FOR HOURLY AVERAGES

Site	χ^2
101	5790
103	875
104	547
116	100
122	88

Supplementing the Chi-Square Program--

It is straightforward to tabulate the highest observed and predicted concentrations for any combination of sites and averaging times. If we are interested in how well the ten highest modeled concentrations agree with the ten highest observed concentrations, a simple tabulation as shown in Figure 18 is easy to interpret. In this example for the three-hour averages, the ten highest observed concentrations were about three times the ten highest predicted concentrations. These highest observed concentrations all occurred at Site 104; most were in late evening or early morning, and all were in late November or December. The ten highest predicted concentrations occurred at four different sites, in the late evening or early morning hours, but at different times of the year (but most often in the late fall/early winter).

```
**********************************************************************

@XQT EVAL.MAXVALUES
 ENTER DATA ELEMENT ADDRESS OR 11 FOR FILE
 SELECTION   1-366,00-23,01-13,1-9,1-9,01-09,1-9,1-9,1-9,1-9
 DATA3HOUR

 10 LARGEST VALUES OF OBSERVED CONCENTRATION WITH DAY, HOUR, AND SITE

  1    2298.   346  20   3  (SITE 104)
  2    2145.   330   2   3  (SITE 104)
  3    1634.   350   8   3  (SITE 104)
  4    1567.   350   2   3  (SITE 104)
  5    1548.   346  23   3  (SITE 104)
  6    1534.   346  17   3  (SITE 104)
  7    1489.   329  23   3  (SITE 104)
  8    1482.   362  14   3  (SITE 104)
  9    1476.   350  23   3  (SITE 104)
 10    1474.   350  20   3  (SITE 104)

 10 LARGEST VALUES OF PREDICTED CONCENTRATION WITH DAY, HOUR, AND SITE

  1     689.   283   2   3  (SITE 104)
  2     681.   296   5   3  (SITE 104)
  3     680.   275   2   3  (SITE 104)
  4     651.   153   2   9  (SITE 115)
  5     617.   198  23   9  (SITE 115)
  6     599.    33  20   1  (SITE 101)
  7     586.   329   2   1  (SITE 101)
  8     575.   275   5   3  (SITE 104)
  9     567.    38   5   1  (SITE 101)
 10     561.   319  23   4  (SITE 105)
```

Figure 18. Computer tables of three-hour averaged ten highest observed and predicted concentrations — all sites.

We recommend the use of tables of extreme values in addition to the chi-square statistic (and displays) in assessing the performance of a model at the high concentrations. If additional computer expense can be tolerated, we further recommend replacing the linear display with the log-frequency displays (Figure 5) while retaining the chi-square computation.

SECTION 7

SOFTWARE

OVERVIEW

This section contains a brief overview of changes made to the original software as documented by Ruff et al. (1980). Not much detail is provided here because most of the basic program logic and FORTRAN IV code in the above reference still applies. Some minor logic and coding changes were made to accommodate differences between the actual RAM/RAPS data base described in Section 2 and the test data base used in our earlier work. Other changes were made to enhance the efficiency of certain programs and to improve several of the output displays. Additionally, a retrieval program was developed to access desired subsets of the data base.

Because the program and data base listings consume an immense paper volume, we have chosen to provide them using two microfiche files. The first, entitled "1976 STL SO_2 RAM HOURLY DATA BASE," contains the hour-by-hour values for predicted and observed SO_2 concentrations, meteorological model input parameters, and summary emissions by area and point source totals per hour. The second microfiche, entitled "1976 STL SO_2 RAM EVALUATION PROGRAM LISTINGS," contains listings of the FORTRAN IV or SPSS codes for the programs and accompanying run streams. The two microfiche display the equivalent of 150 and 180 pages of computer listings respectively. A brief discussion of the information contained on each microfiche is presented below.

FILE DESCRIPTIONS

In this section, terminology is used specific to the EPA UNIVAC 1110 system. Understanding of these terms presupposes a knowledge of this system or a referral to our previous report (Ruff et al., 1980).

The information given on the microfiche covers listings of four different files developed on EPA's Univac 1110 computer system at Research Triangle Park in the course of the RAM evaluation. The hourly data base (1976 STL SO_2 RAM HOURLY DATA BASE) is contained in element DATA1 of the file SRI*D.DATA1; the program listings (1976 STL SO_2 RAM EVALUATION PROGRAM LISTINGS) are contained on three files: SRI*SRI., SRI*EVAL., and SRI*RON.. (In future references, the "SRI*" is omitted when referring to files.)

The file element--SRI*D.DATA1--contains the hourly averaged data. There are 8784 lines of data, one for each hour. Each line is 132 characters long. The format is given in Table 7. The minimum value for observed concentrations

TABLE 7. FORMAT FOR THE 1976 HOURLY DATA BASE

Columns	Format*	Parameter and Range (or Units)
1 - 3	I3	Julian day (1-366)
4 - 5	I2	Hour (00-23)
6 - 7	I2	Pasquill-Gifford stability class 1-7 (A-G)
8 - 10	I3	Wind speed (m/s) × 10
11 - 13	I3	Wind direction (degrees)
14 - 16	I3	Temperature (°C)
17 - 20	I4	Total point emissions (g/s) ÷ 100
21 - 24	I4	Total area emissions (g/s)
25 - 28	I4	Mixing height (m)
29 - 80	13I4	Observed concentrations ($\mu g/m^3$) for Sites 101, 103, 104, 105, 106, 108, 113, 114, 115, 116, 120, 121, and 122 respectively.
81 - 132	13I4	Predicted concentrations ($\mu g/m^3$) for Sites 101, 103, 104, 105, 106, 108, 113, 114, 115, 116, 120, 121, and 122 respectively.

*As it would be specified in a Fortran FORMAT statement

is 6 $\mu g/m^3$ because the instrument did not have sufficient accuracy to represent concentrations below this number. Predicted concentrations were allowed to assume values down to zero. Missing data in the file are marked with a "-1."

The first program file, SRI., contains the source main programs and source subroutines of the analysis programs:

REGANA

CHIFIT

FREQ

ACSCOR

COREL2

TIMSER

Corresponding subroutines have version names attached to them in most cases as follows:

ACSCOR/STAT01

TIMSER/STAT02

CHIFIT/STAT03

REGANA/STAT04

COREL2/STAT05

Additional elements are included to aid in mapping the analysis programs. The map elements and their corresponding programs are:

MAPFREQ	FREQ
MAPIT/STAT01	ACSCOR
MAPIT/STAT02	TIMSER
MAPIT/STAT03	CHIFIT
MAPIT/STAT04	REGANA
MAPIT/STAT05	COREL2

The file EVAL. contains all the absolute elements that were saved, as well as several special-purpose programs and subroutines. In general, these special programs are not intended to be user programs and do not represent statistical or any of the associated graphical code. Their primary purpose, as briefly described by "comment" statements within the code, is to manipulate the data base before executing one of the analysis programs in files SRI. and RON.. One program, EVAL. RAMDATARET, was particularly useful in retrieving user-selected subsets from SRI*D.DATA1. The ability to manipulate and retrieve data interactively is essential in the model evaluation process. The example interactive session, shown in Figure 19, demonstrates how RAMDATARET allows a user to select and retrieve a data set based on desired parameter ranges.

The SPSS programs are stored in file RON., which contains the code for data format, discriminant analysis, and multiple regression procedures. The element names have prefixes of "DAT," "DIS," and "REG" for programs used to format data into SPSS compatible files, conduct discriminant analyses, and conduct multiple regression analyses, respectively.

```
@XQT EVAL.RAMDATARET                    >
 RAM MODEL EVALUATION DATA RETRIEVAL PROCEDURE
 THERE ARE 307440 VALUES IN THE DATA BASE
 DAY, HOUR, 1(7) ATMOSPHERIC STABILITY, 2(7) WIND SPEED, 3(16) WIND DIRECT
ION,
 4(5) TEMPERATURE, 5(4) POINT EMISSIONS, 6(4) AREA EMISSIONS,
 7(6) MIXING HEIGHT

 ENTER START AND END DAY, RANGE 1 TO 366, EG 32 60
>1 366
 ENTER START AND END HOUR, RANGE 0 TO 23, EG 0 23
>0 23
 ENTER SELECTION PARAMETER NUMBER OR 0 TO CONTINUE, EG 4
>0
 ENTER THE STATION RANGE WHERE 1=101,2=103,3=104,4=105,5=106,6=108,
 7=113,8=114,9=115,10=116,11=120,12=121,13=122, EG 2 5
>1 1

 YOU HAVE REQUESTED UP TO  17568 CONCENTRATION VALUES
 PARAMETER SELECTION RANGES
 PARAMETER    1   2   3   4   5   6   7   8   9  10
 MINIMUM      1   0   1   1   1   1   1   1   1   1
 MAXIMUM    366  23   1   7   7  16   5   4   4   6
 OUTPUT WRITTEN TO UNIT 11, DATE, TIME, OBSERVED, PREDICTED VALUES
 ENTER 6 IF YOU WANT PRINTED OUTPUT ALSO
>0
```

Figure 19. Example of a data retrieval session using RAMDATARET.

REFERENCES

Bowne, N. E., 1980: "Validation and Performance Criteria Air Quality Models" Conference Papers, Second Joint Conference on Applications of Air Pollution Meteorology, AMS.

Draxler, R. R., 1979: "An Improved Gaussian Model for Long-Term Average Air Concentration Estimates," NOAA Air Resources Laboratories, Silver Spring, MD 20910 (unpublished manuscript).

Guldberg, P. H., and C. W. Kern, 1978: "A Comparative Validation of the RAM and PTMTP Models for Short-Term SO_2 Concentrations in Two Urban Areas," JAPCA, Vol. 28, No. 9.

Hayes, S. R., 1979: "Performance Measures and Standards for Air Quality Simulation Models," Final Report, EPA-450/4-79-032, 1100 pp.

Hribar, J. R., 1976: "EPA Quality Assurance Audits," EPA Report No. EPA-600/4-76-032.

Irwin, J., 1980: Personal Communications.

Javitz, H. S., and R. E. Ruff, 1980: "Evaluation of the Real-Time Air-Quality Model Using the RAPS Data Base; Volume 2. Statistical Procedures," EPA Report No. EPA-600/4-80-013b.

Klein, O. C., and F. E. Littman, 1978: "Quality Assurance Audits," Final Report, Task Order No. 106, EPA Contract 68-02-2093.

Martinez, J. R., H. S. Javitz, R. E. Ruff, and W. F. Dabberdt, 1980: "Development and Application of Methods for Evaluating Highway Air Pollution Dispersion Models," Conference Papers, Second Joint Conference on Applications of Air Pollution Meteorology, AMS.

Nie, N. H., C. H. Hull, J. G. Jenkins, K. Steinbrenner, and D. H. Bent, 1975: Statistical Package for the Social Sciences (McGraw-Hill Book Co., New York, New York).

Ruff, R. E., 1980: "Evaluation of the Real-Time Air-Quality Model Using the RAPS Data Base; Volume 1. Overview," EPA Report No. EPA-600/4-80-013a.

Ruff, R. E., 1980: "Evaluation of the Real-Time Air-Quality Model Using the RAPS Data Base; Volume 4. Evaluation Guide," EPA Report No. EPA-600/4-80-013d.

Ruff, R. E., H. Shigeishi, and R. H. Allen, 1980: "Evaluation of the Real-Time Air-Quality Model Using the RAPS Data Base; Volume 3. Program User's Guide," EPA Report No. EPA-600/4-80-013c.

Schiermeier, F. A., 1978: "Air Monitoring Milestones: RAPS Field Measurements are in," Env. Sci. & Tech., Vol. 12, No. 6.

Turner, D. B., 1970: Workbook of Atmospheric Dispersion Estimates, Office of Air Programs Publication, No. AP-26, EPA, Research Triangle Park, NC.

Turner, D. B., and J. H. Novak, 1978: "Users Guide for RAM," EPA-600/8-78-016 a & b, Environmental Protection Agency Publication (2 Volumes) (1978).

Venkatram, A., 1979: "A Note on the Measurement and Modeling of Pollutant Concentrations Associated with Point Sources," Boundary-Layer Meteorology, 17, 523-526.

APPENDIX A

FREQUENCY DISTRIBUTION OF RAPS HOURLY METEOROLOGICAL PARAMETERS

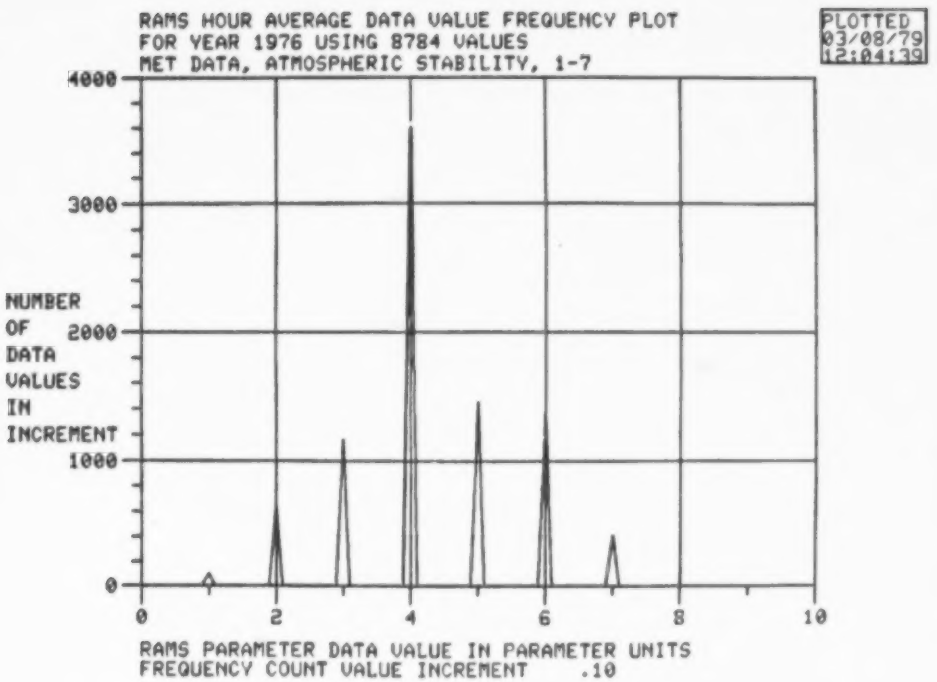


Figure A-1. Frequency distribution of atmospheric stability classes for 1976.

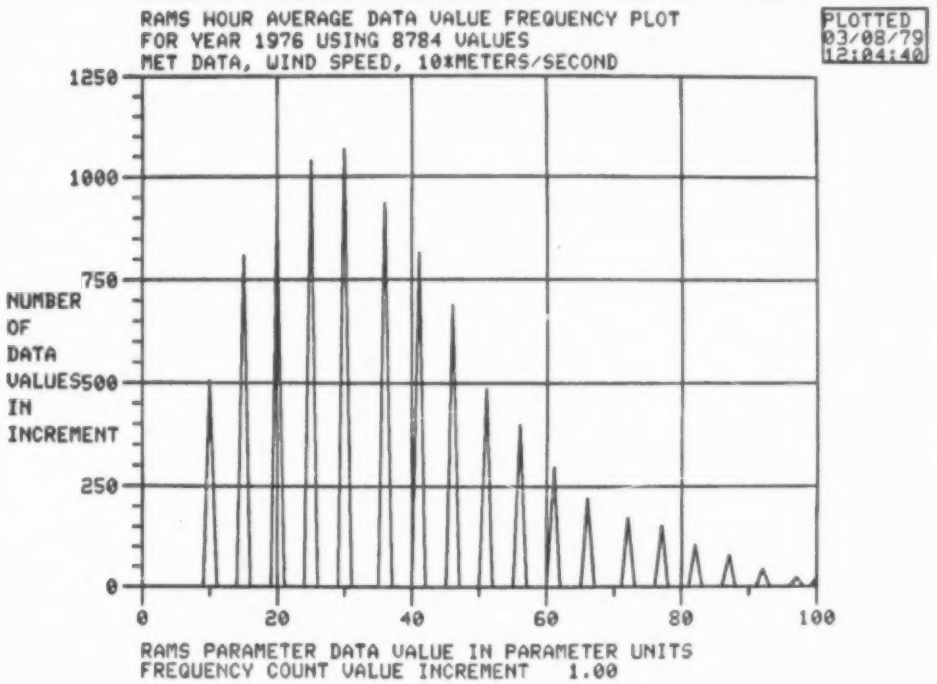


Figure A-2. Distribution of wind speed for 1976.

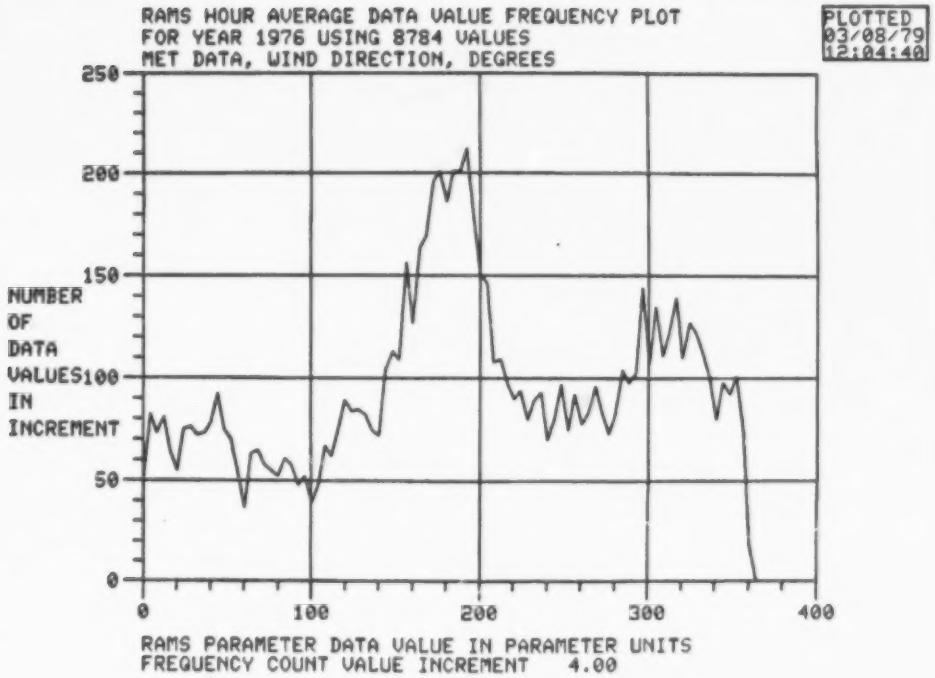


Figure A-3. Frequency distribution of wind direction for 1976.

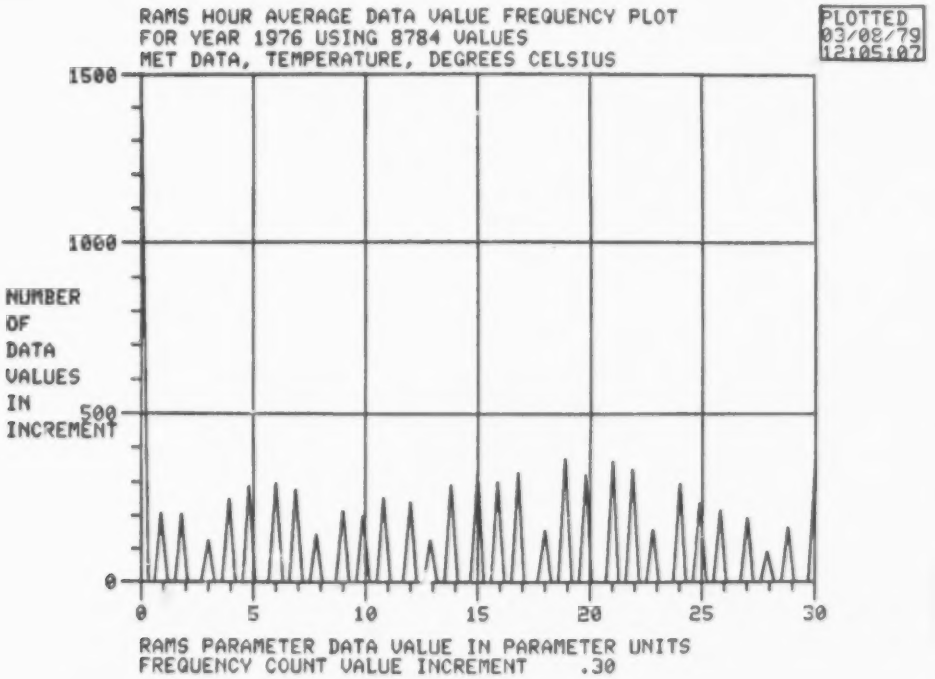


Figure A-4. Frequency distribution of temperature for 1976.

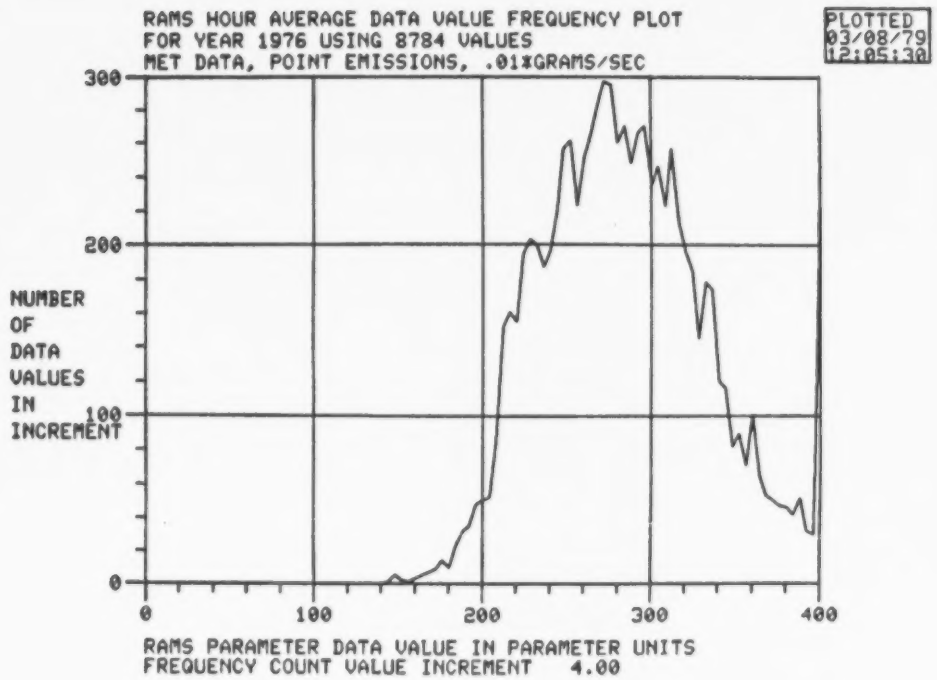


Figure A-5. Frequency distribution of point source emissions for 1976.

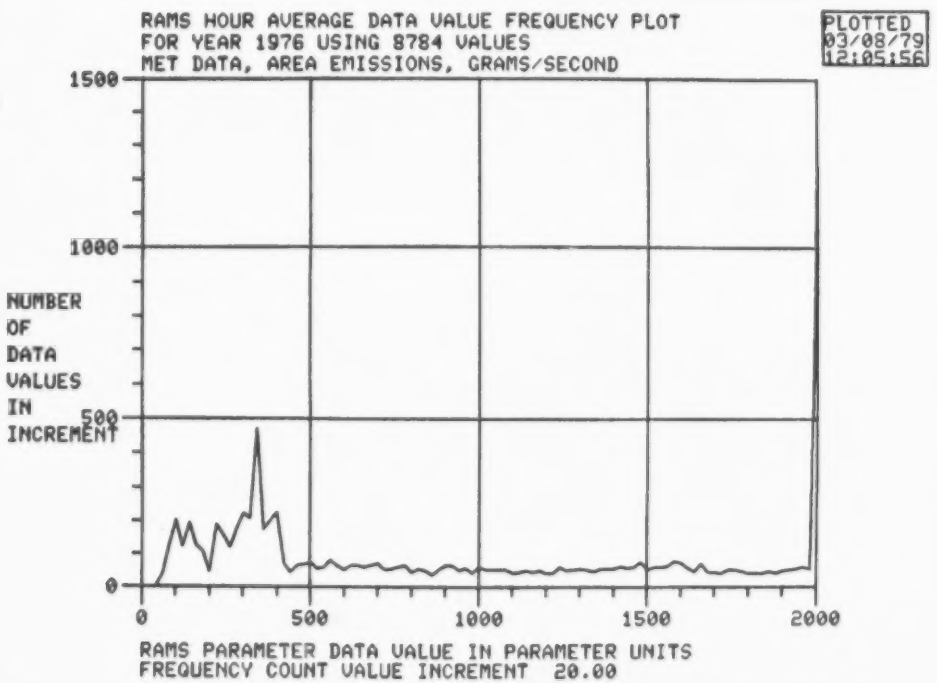


Figure A-6. Frequency distribution of area source emissions for 1976.

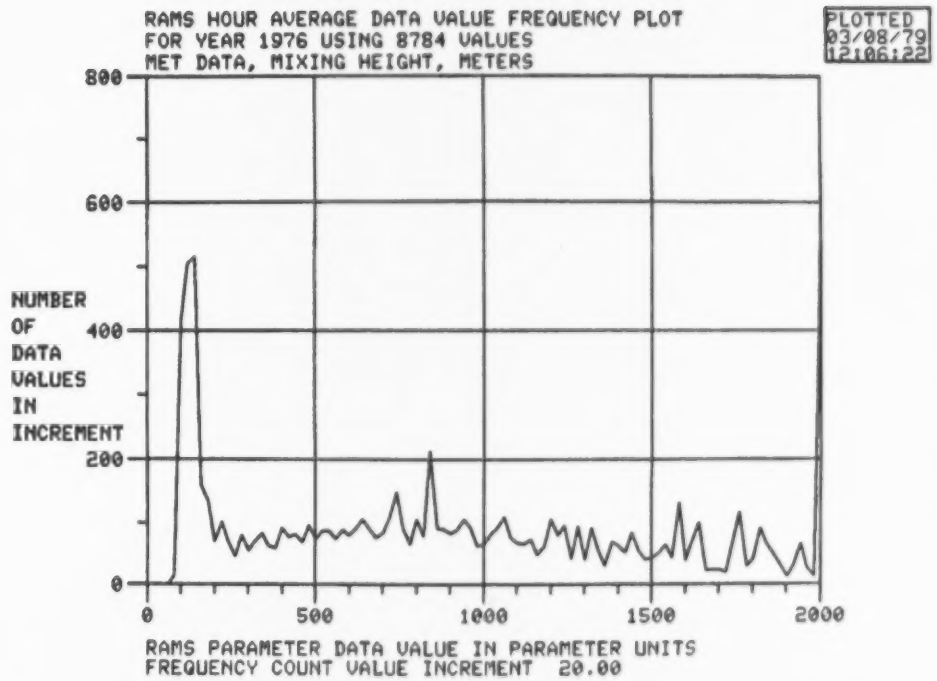


Figure A-7. Frequency distribution of mixing height for 1976.

APPENDIX B

LOGARITHM-FREQUENCY GRAPHS OF OBSERVED
AND PREDICTED CONCENTRATIONS
(Hourly averages $\mu g/m^3$)

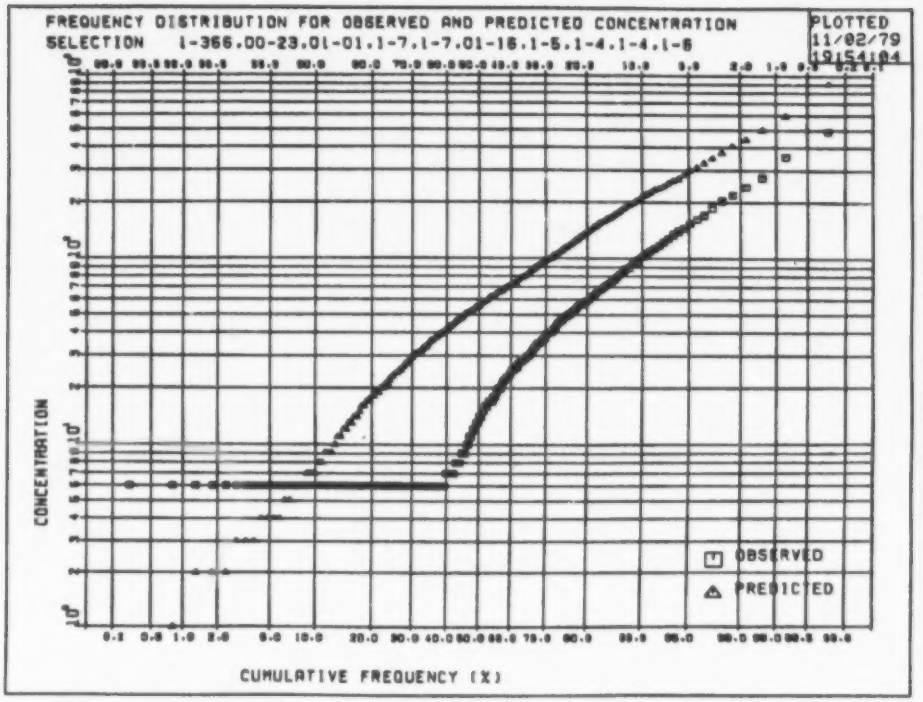


Figure B-1. Observed and predicted concentrations for Station 101.

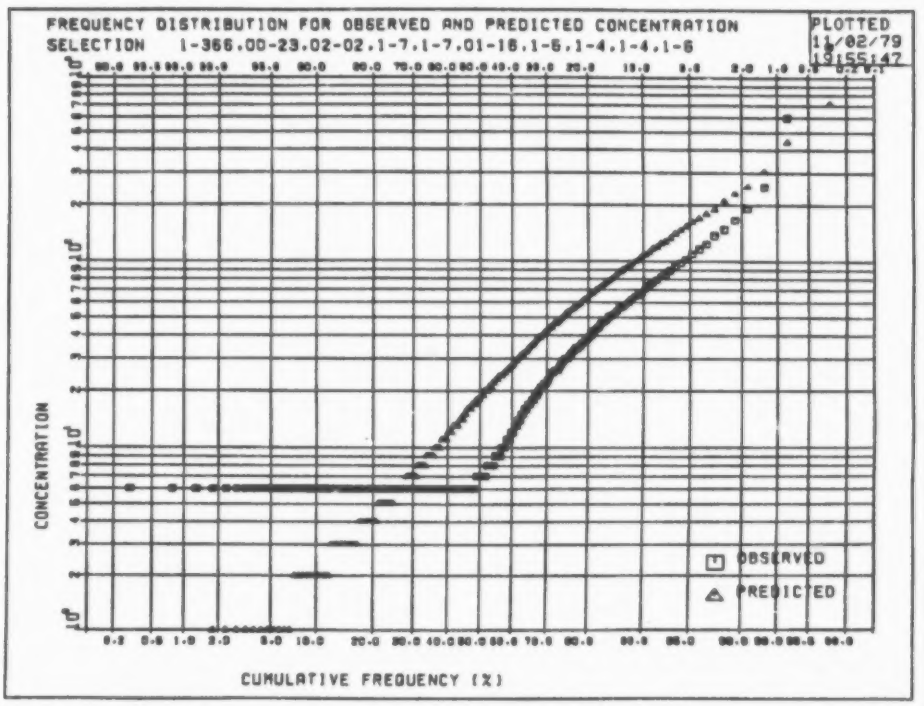


Figure B-2. Observed and predicted concentrations for Station 103.

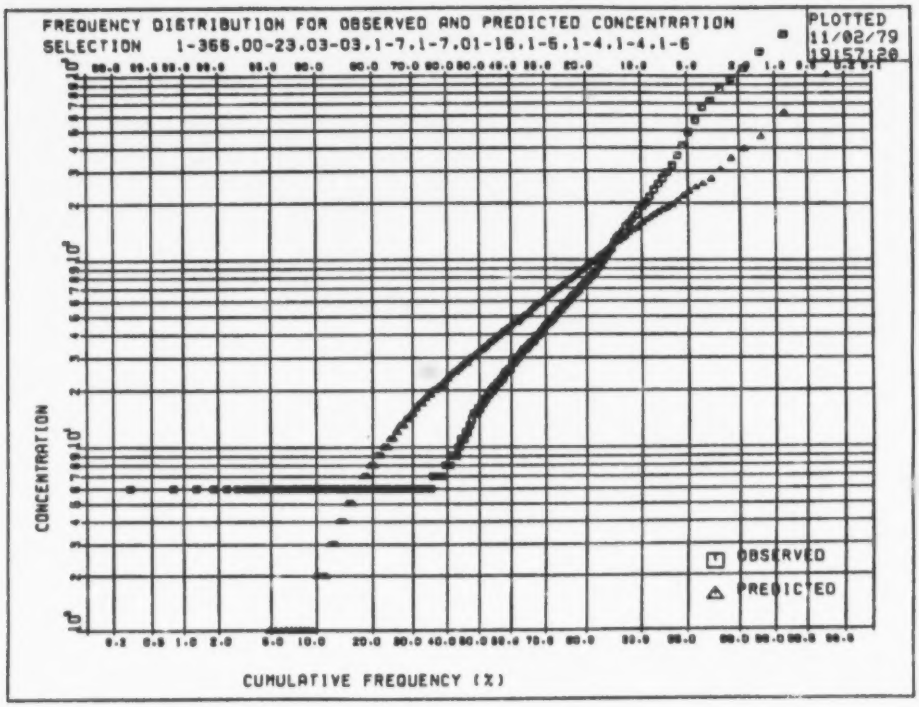


Figure B-3. Observed and predicted concentrations for Station 104.

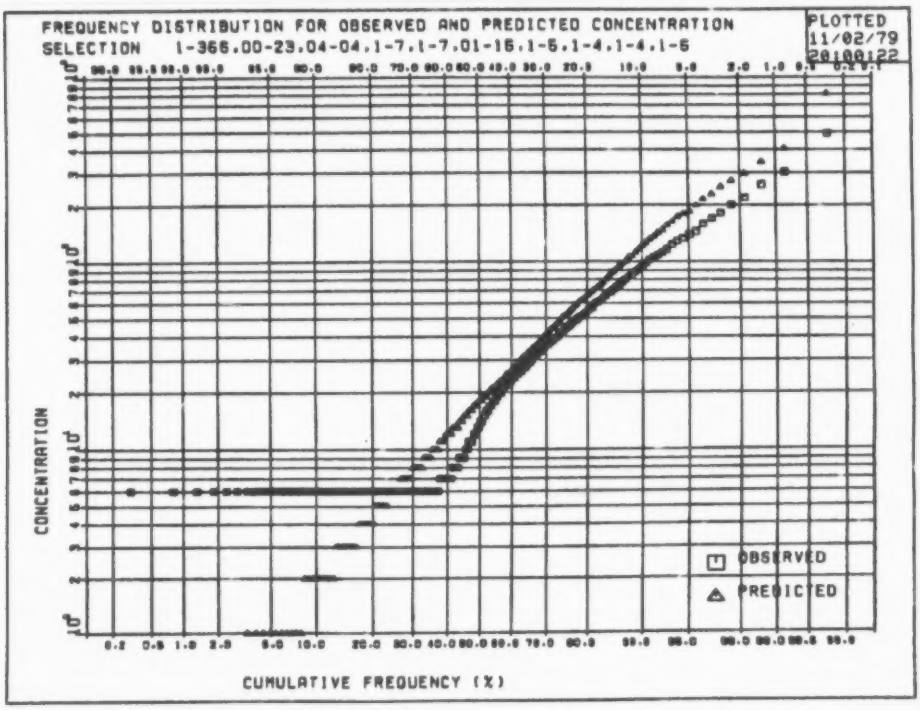


Figure B-4. Observed and predicted concentrations for Station 105.

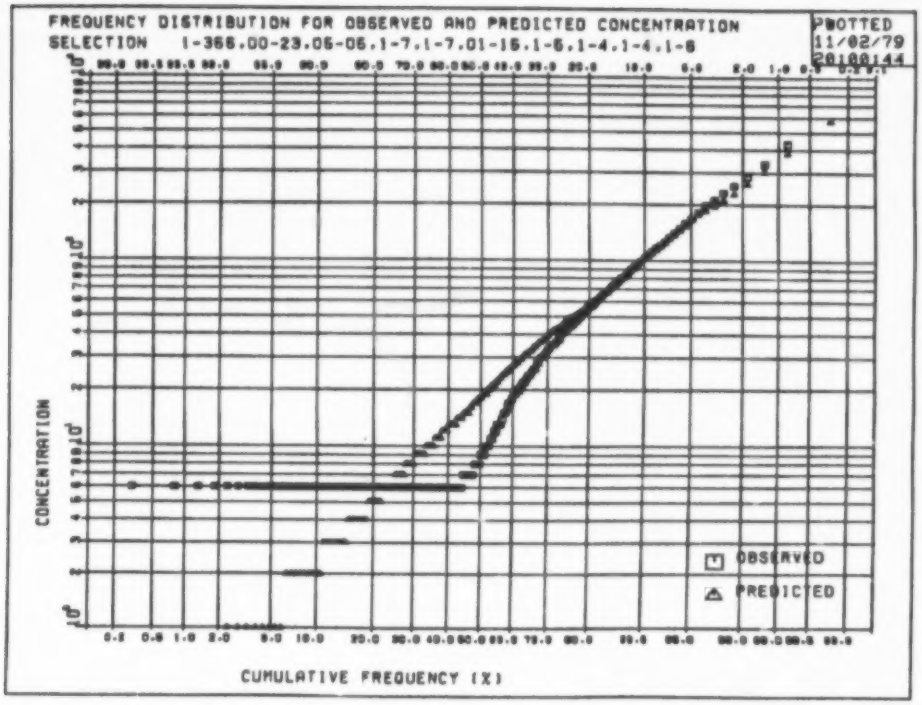


Figure B-5. Observed and predicted concentrations for Station 106.

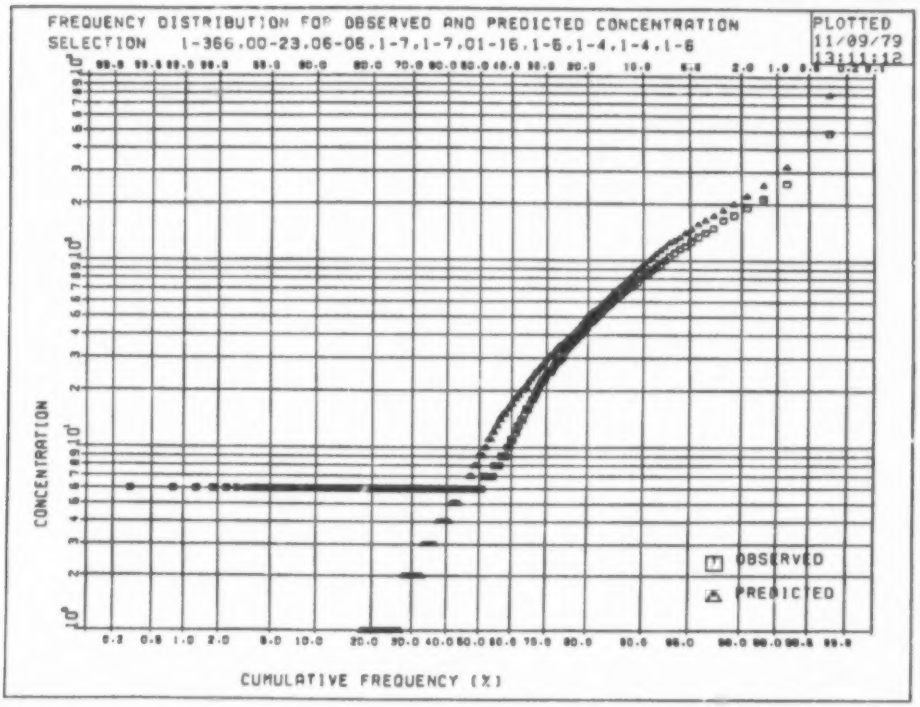


Figure B-6. Observed and predicted concentrations for Station 108.

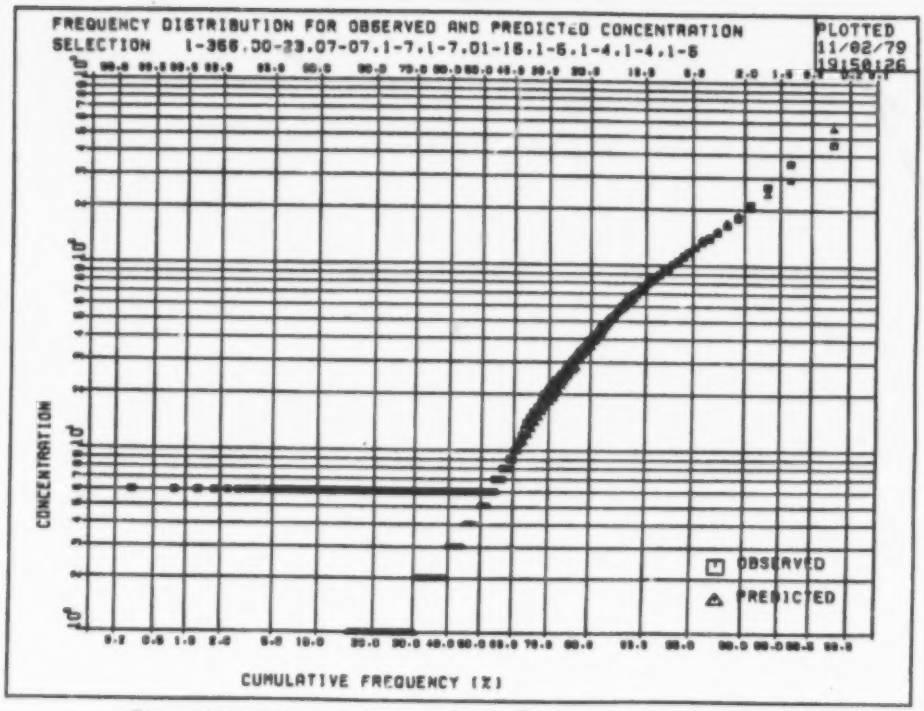


Figure B-7. Observed and predicted concentrations for Station 113.

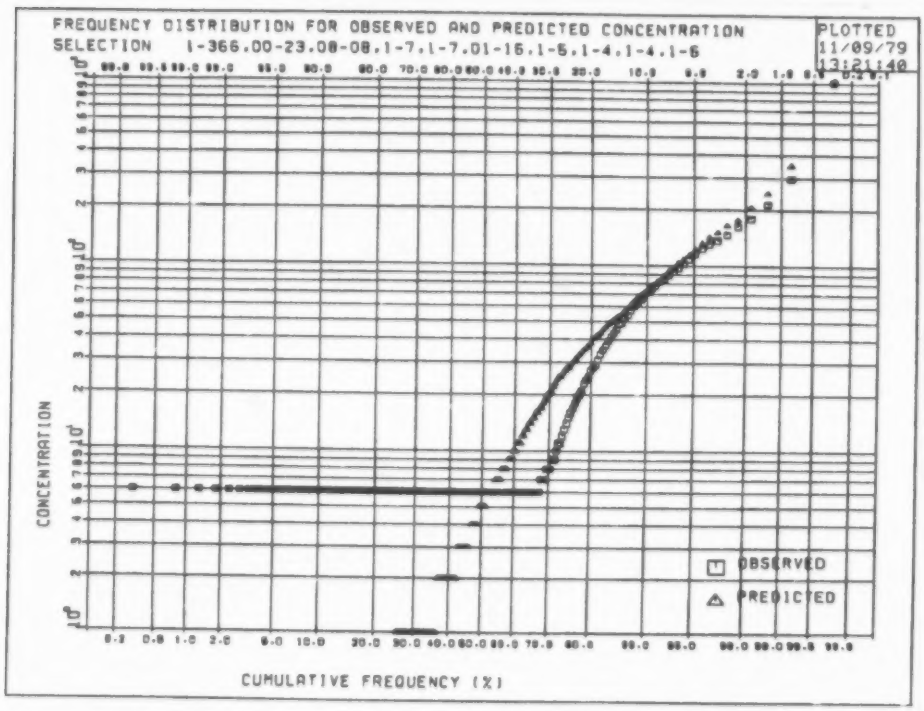


Figure B-8. Observed and predicted concentrations for Station 114.

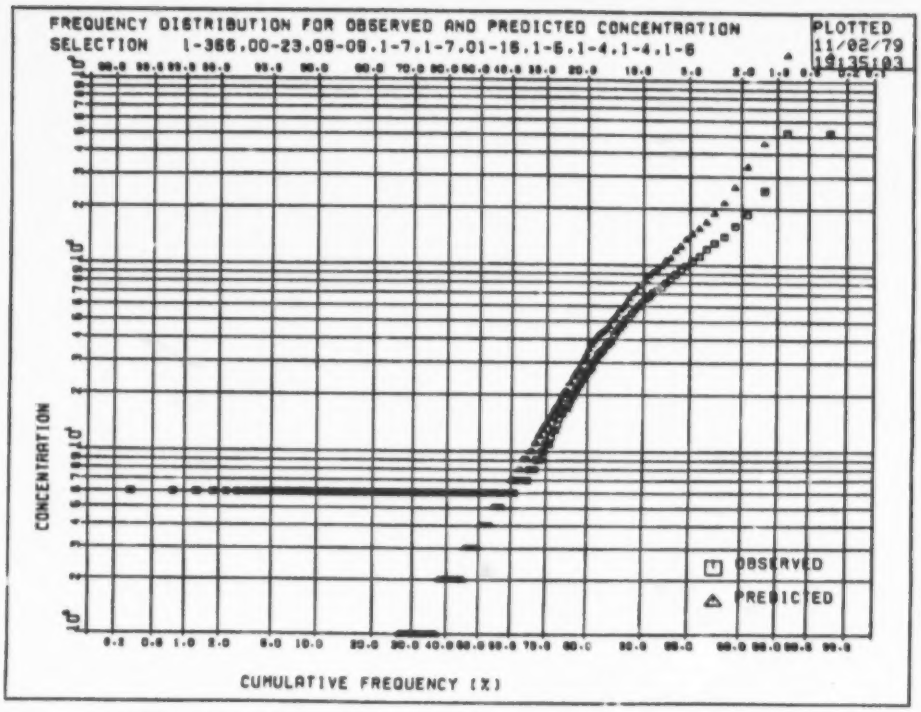


Figure B-9. Observed and predicted concentrations for Station 115.

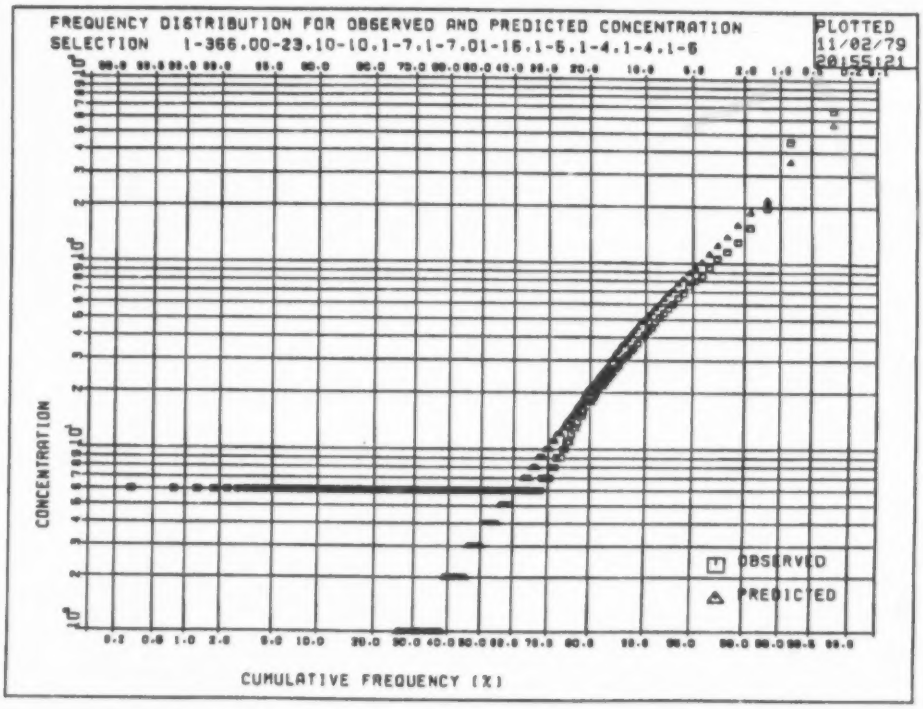


Figure B-10. Observed and predicted concentrations for Station 116.

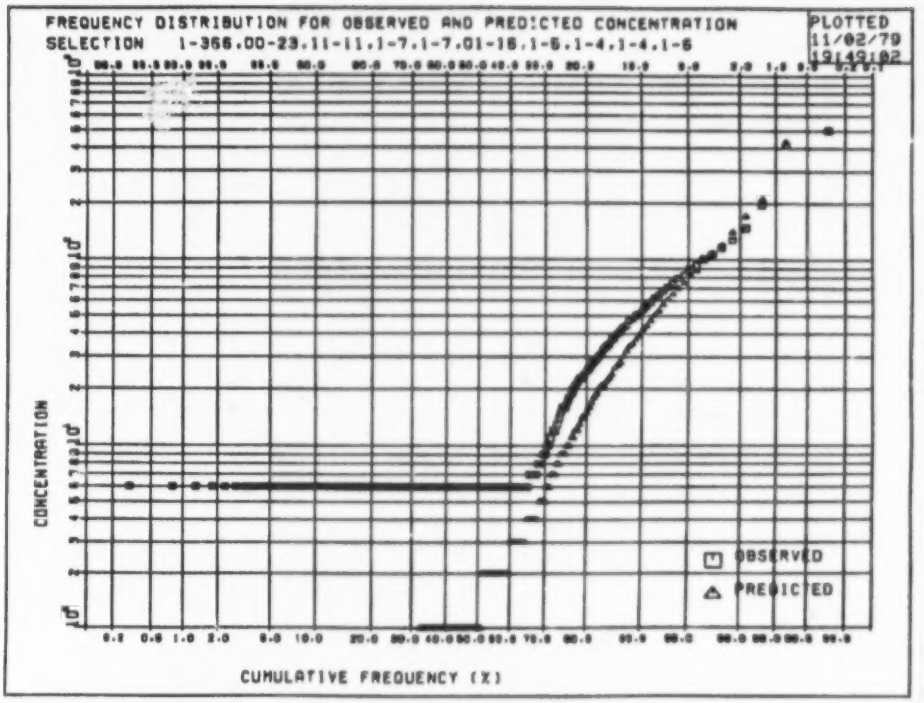


Figure B-11. Observed and predicted concentrations for Station 120.

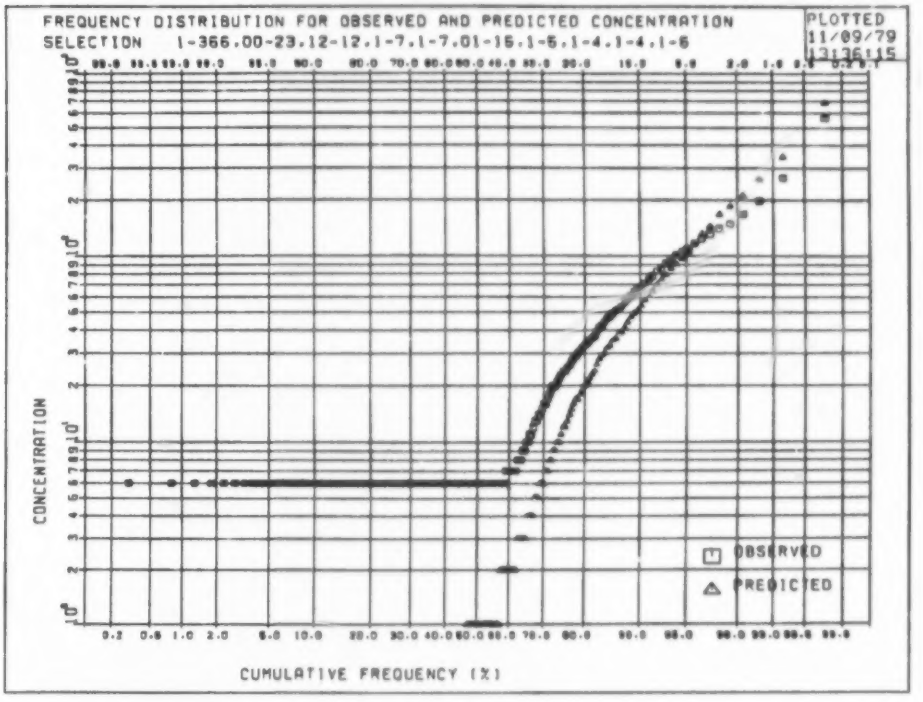


Figure B-12. Observed and predicted concentrations for Station 121.

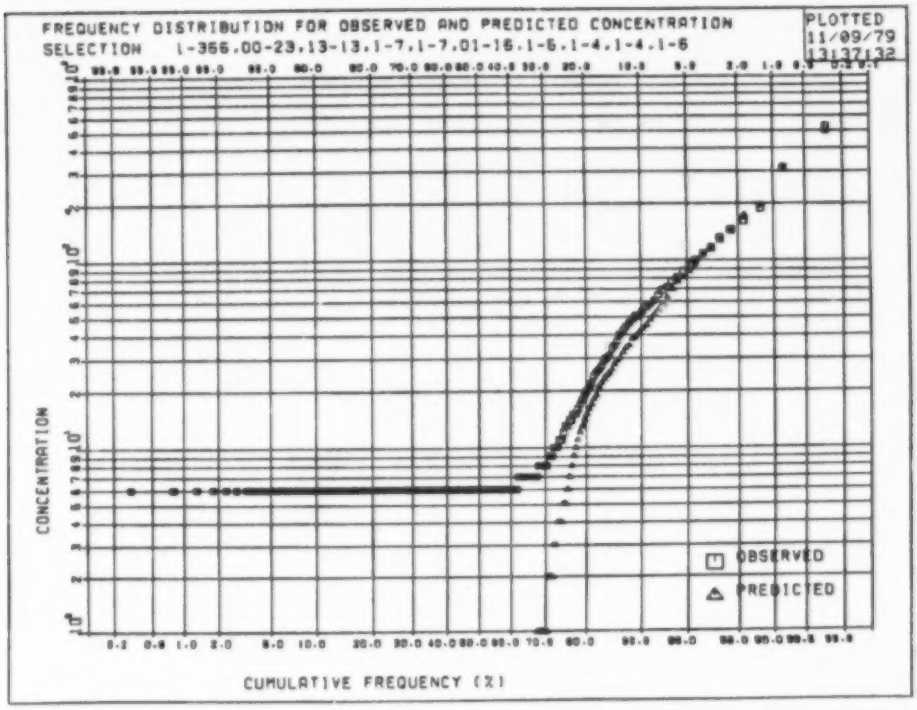


Figure B-13. Observed and predicted concentrations for Station 122.

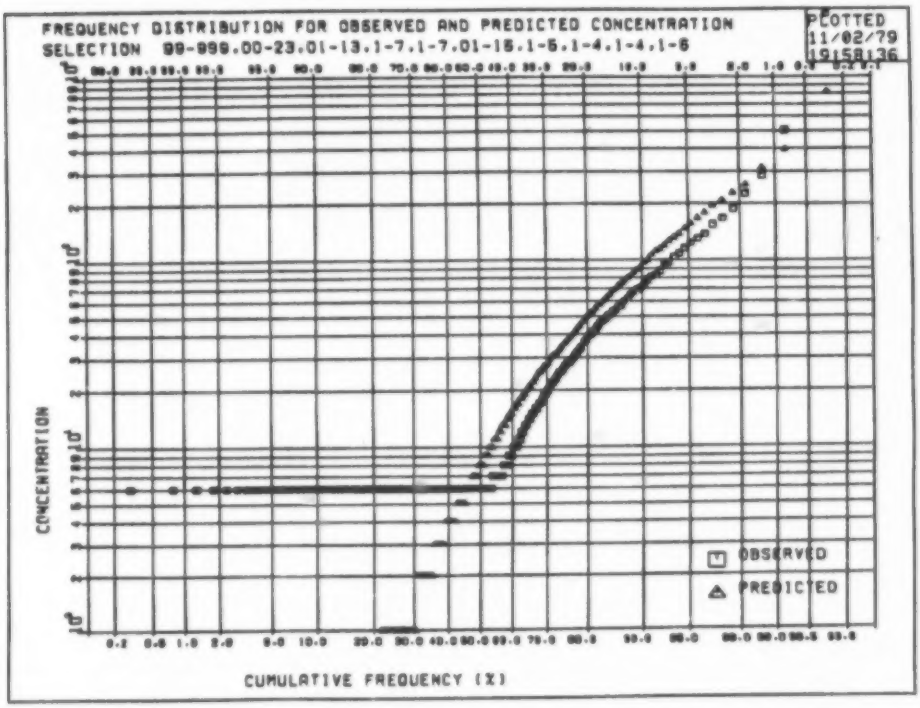


Figure B-14. Composite observed and predicted concentrations for all stations.

APPENDIX C

SCATTERPLOTS OF HOURLY PREDICTED VERSUS OBSERVED SO_2 CONCENTRATIONS FOR SELECTED RAMS MONITORING SITES

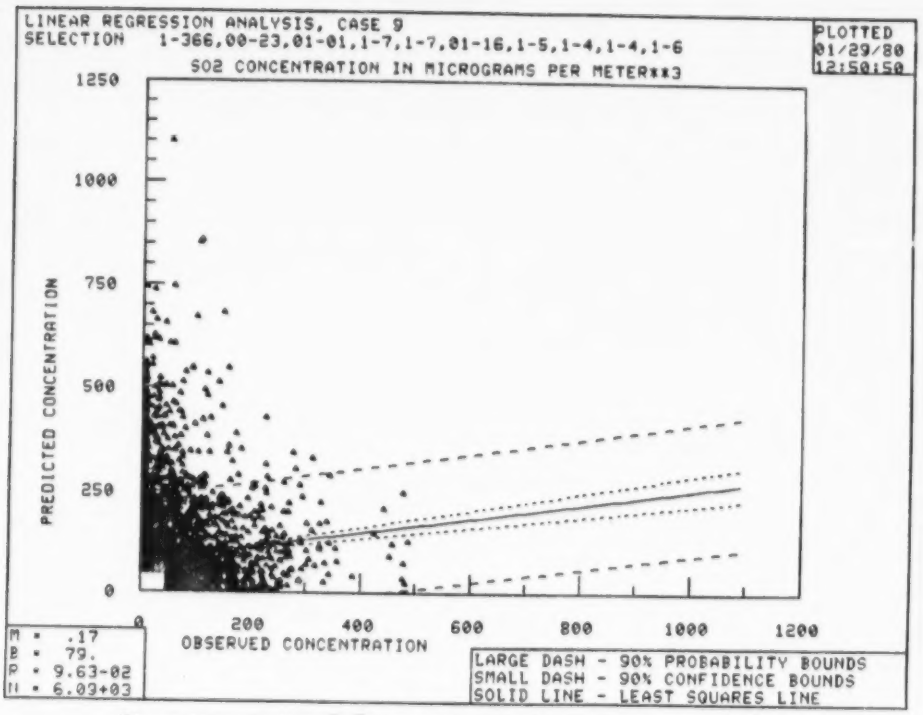


Figure C-1. Scatterplot of all 1976 hourly SO_2 averages for Site 101.

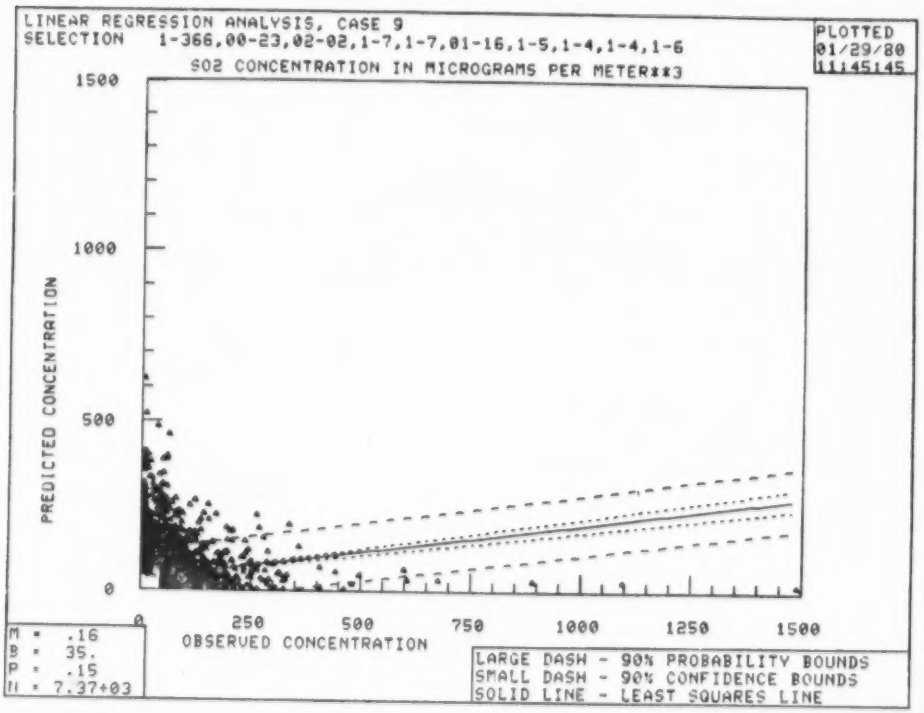


Figure C-2. Scatterplot of all 1976 hourly SO_2 averages for Site 103.

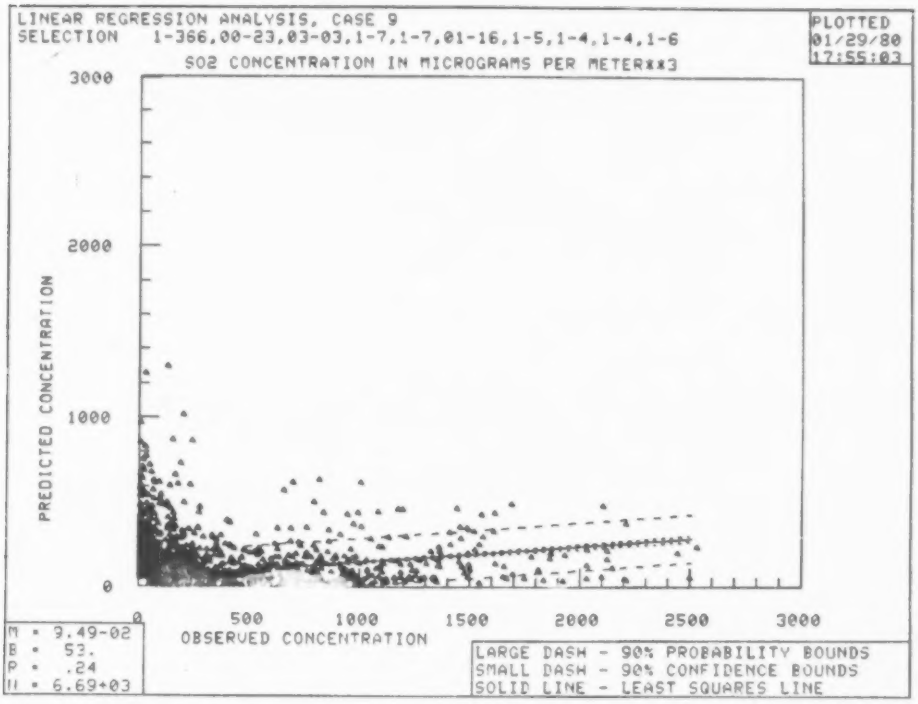


Figure C-3. Scatterplot of all 1976 hourly SO_2 averages for Site 104.

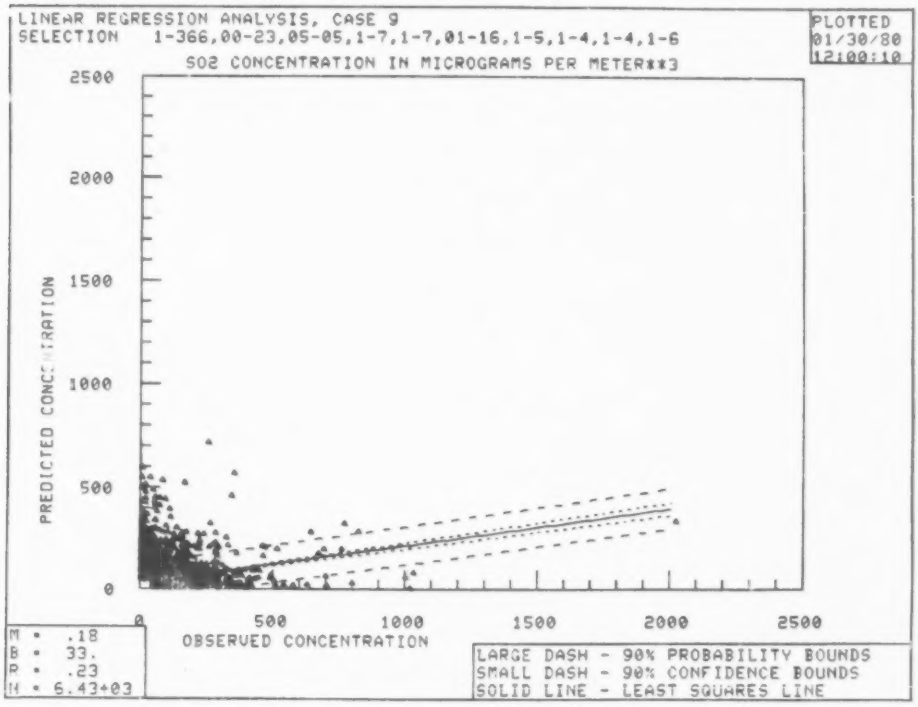


Figure C-4. Scatterplot of all 1976 hourly SO_2 averages for Site 106.

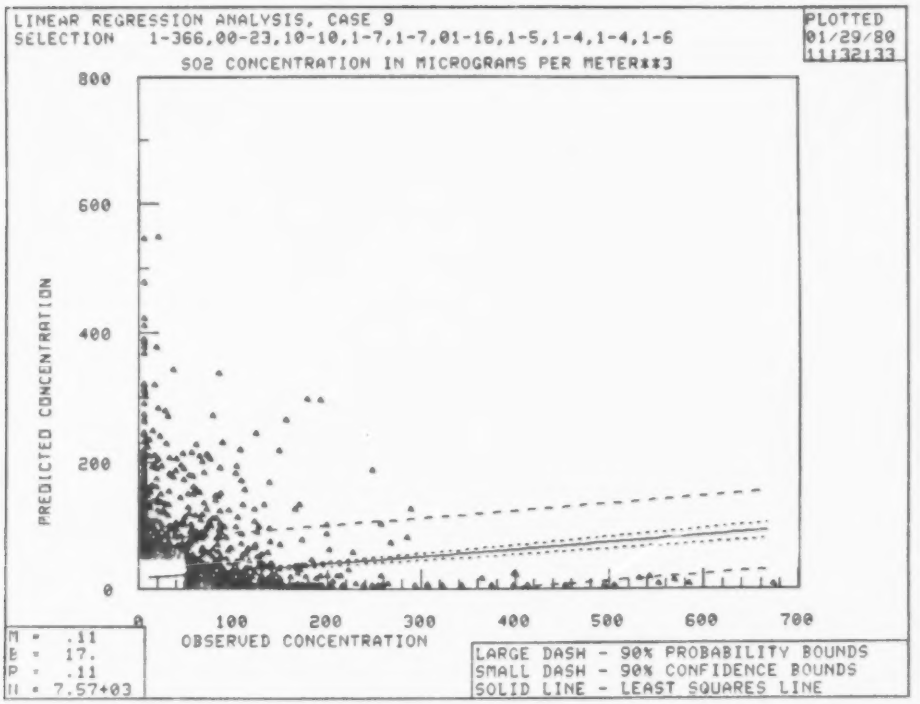


Figure C-5. Scatterplot of all 1976 hourly SO_2 averages for Site 116.

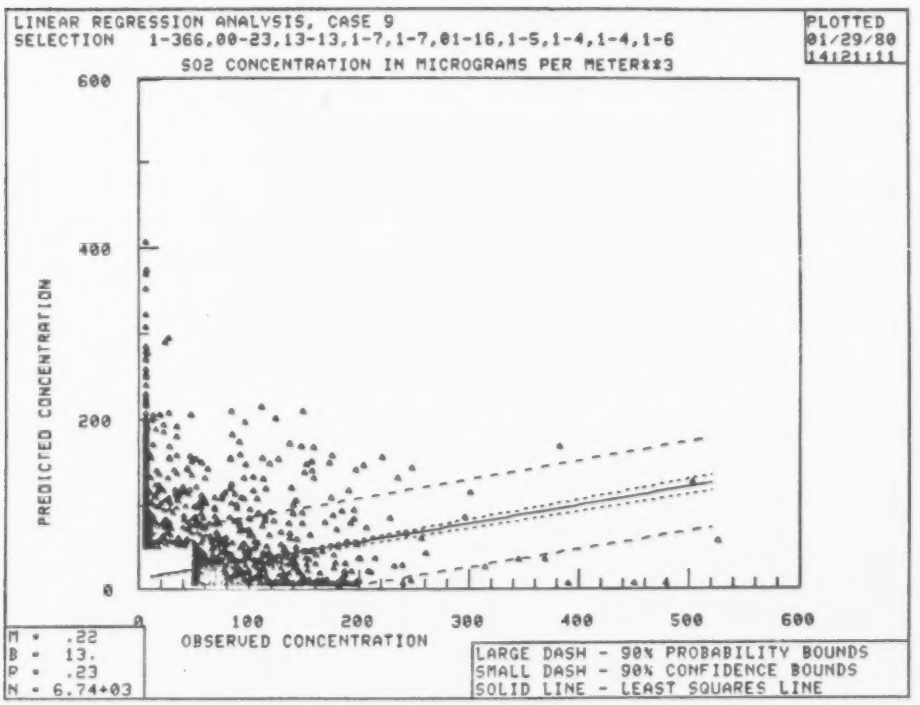


Figure C-6. Scatterplot of all 1976 hourly SO_2 averages for Site 122.

APPENDIX D

IRWIN'S RESULTS FOR SITE 104

Bottom of 8½ x 11 paper

TABLE D-1. SUMMARY OF RESIDUALS FOR JAN-FEB-MAR 1976

Wind Direction	Above 1000	1000 to 800	800 to 600	600 to 400	400 to 200	+200 to -200	-200 to -400	-400 to -600	-600 to -800	Above -800
0-20						6	2			
21-40						15				
41-60						10	2			
61-80						17				
81-100						11				
101-120						7				
121-140					1	20				
141-160						11	2			
161-180						22				
181-200						30	4			
201-220						17	3			
221-240						16				
241-260						24	1			
261-280					1	54	3			
281-300						157	1			
301-320						78	3			
321-340						36	1			
341-360						20	1			
					2	551	23			

Notes: Totals represent the number of occurrences at Site 104 within the indicated residual class.

Residual = Observed Concentration - Predicted Concentration.

TABLE D-2. SUMMARY OF RESIDUALS FOR APR-MAY-JUN 1976

Wind Direction	Above 1000	1000 to 800	800 to 600	600 to 400	400 to 200	+200 to -200	-200 to -400	-400 to -600	-600 to -800	Above -800
0-20						25				
21-40						33				
41-60						30				
61-80						34				
81-100						17				
101-120						29				
121-140						37				
141-160						11				
161-180						39				
181-200						32	1			
201-220						17	3	1	1	1
221-240					3	20		1		
241-260						33				
261-280						22	1			
281-300						28				
301-320						39				
321-340						17				
341-360						19				
					3	482	5	2	1	1

Notes: Totals represent the number of occurrences at Site 104 within the indicated residual class.

Residual = Observed Concentration - Predicted Concentration.

TABLE D-3. SUMMARY OF RESIDUALS FOR JUL-AUG-SEP 1976

Wind Direction	Above 1000	1000 to 800	800 to 600	600 to 400	400 to 200	+200 to -200	-200 to -400	-400 to -600	-600 to -800	Above -800
0-20						30	1			
21-40						26				
41-60						29				
61-80						24				
81-100						25				
101-120						32				
121-140						49				
141-160						55	3			
161-180						66				
181-200						42				
201-220						8	7	3	1	2
221-240						9	1	1		
241-260						8				
261-280						27				
281-300						12				
301-320						21				
321-340						30				
341-360						29	1			
						522	13	4	1	2

Notes: Totals represent the number of occurrences at Site 104 within the indicated residual class.

Residual = Observed Concentration - Predicted Concentration.

TABLE D-4. SUMMARY OF RESIDUALS FOR OCT-NOV-DEC 1976

Wind Direction	Above 1000	1000 to 800	800 to 600	600 to 400	400 to 200	+200 to -200	-200 to -400	-400 to -600	-600 to -800	Above -800
0-20					4	34	2			
21-40			1			19				
41-60		2				12	2			
61-80						10				
81-100						6				
101-120			1		1	1				
121-140	1		1	1		10				
141-160	1	1	1		2	14				
161-180						16				
181-200	2		1	3	8	34				
201-220	1			1	4	8	3	4	3	
221-240	3		1	1		13	1			
241-260	3		1	2		21				
261-280	3	1		3	5	42				
281-300						69	6			
301-320			1		1	113	5			
321-340			1	2	3	125				
341-360		1	2		4	82				
	14	5	11	13	32	629	19	4	3	

Notes: Totals represent the number of occurrences at Site 104 within the indicated residual class.

Residual = Observed Concentration - Predicted Concentration.

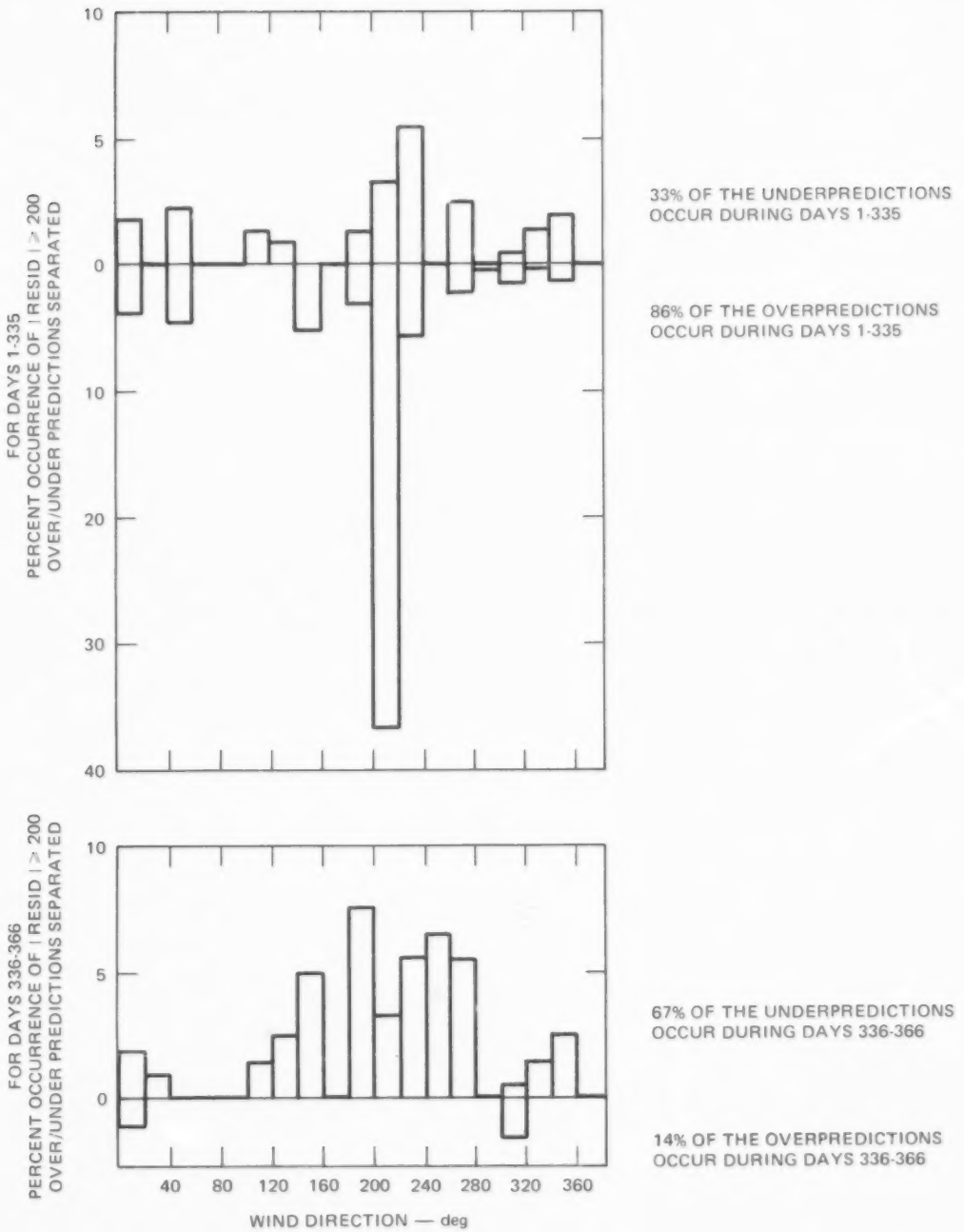


Figure D-1. Distribution of Site 104 residuals by wind direction.

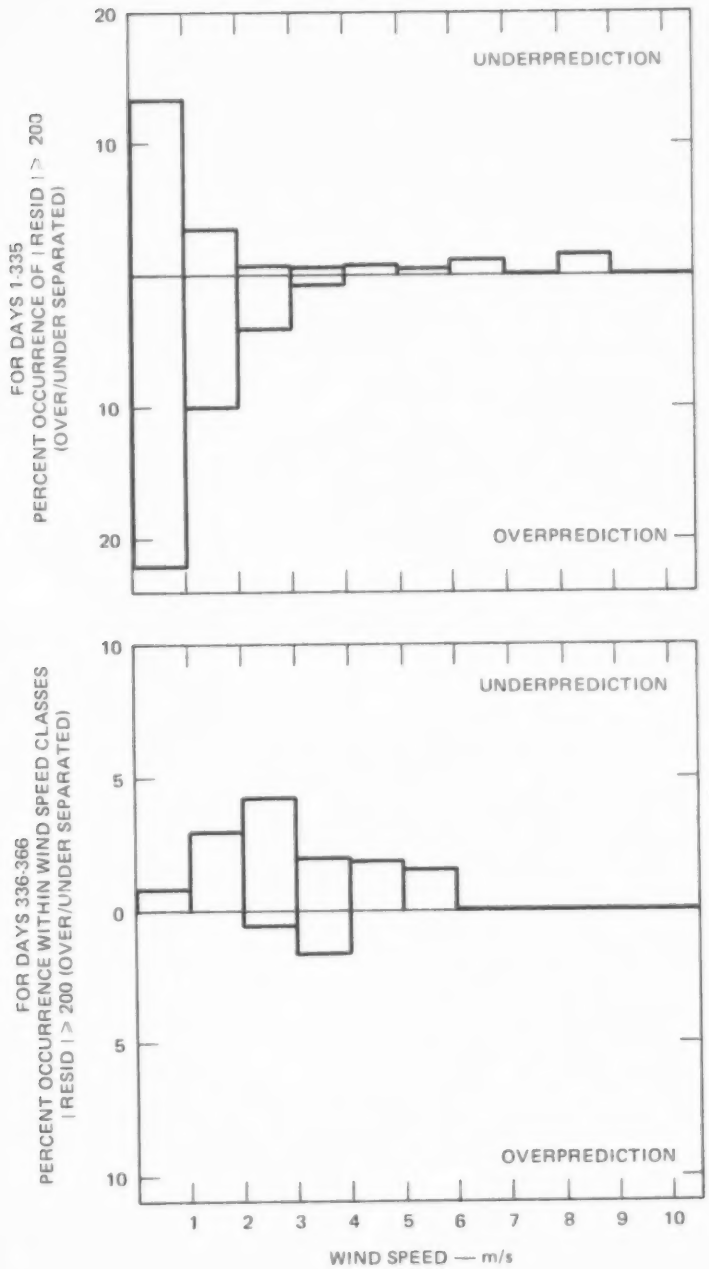


Figure D-2. Distribution of Site 104 residuals by wind speed.

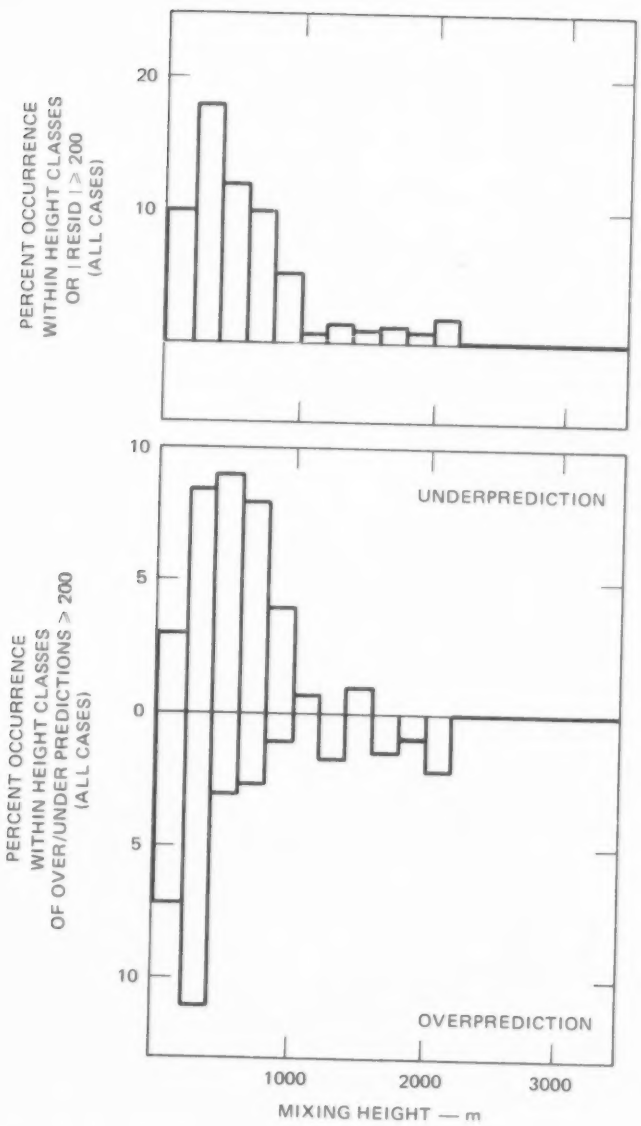


Figure D-3. Distribution of Site 104 residuals by mixing height.

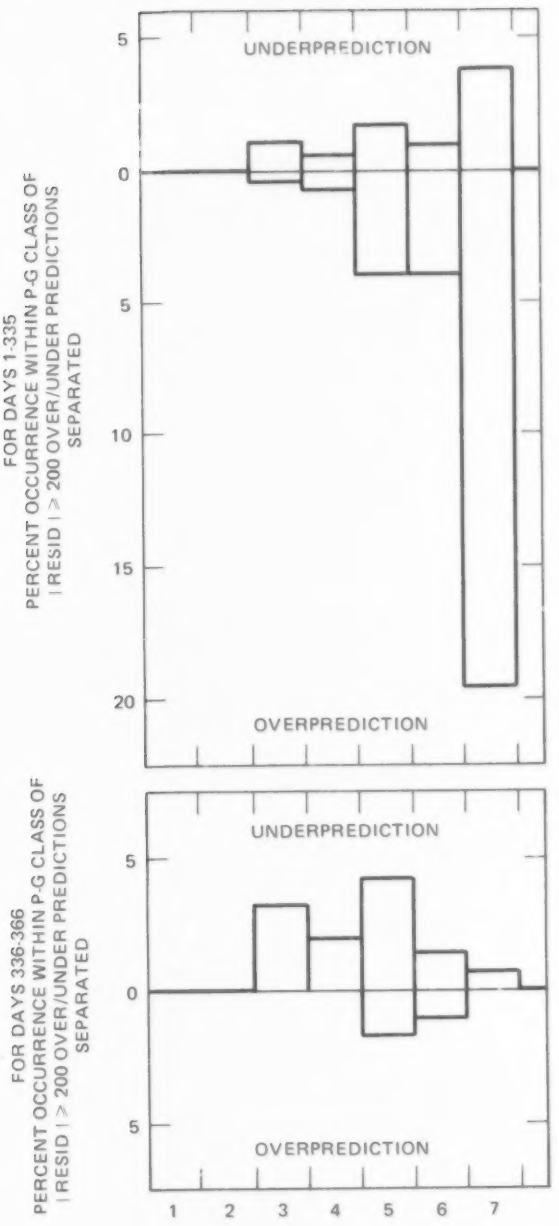


Figure D-4. Distribution of Site 104 residuals by Pasquill-Gifford stability class.

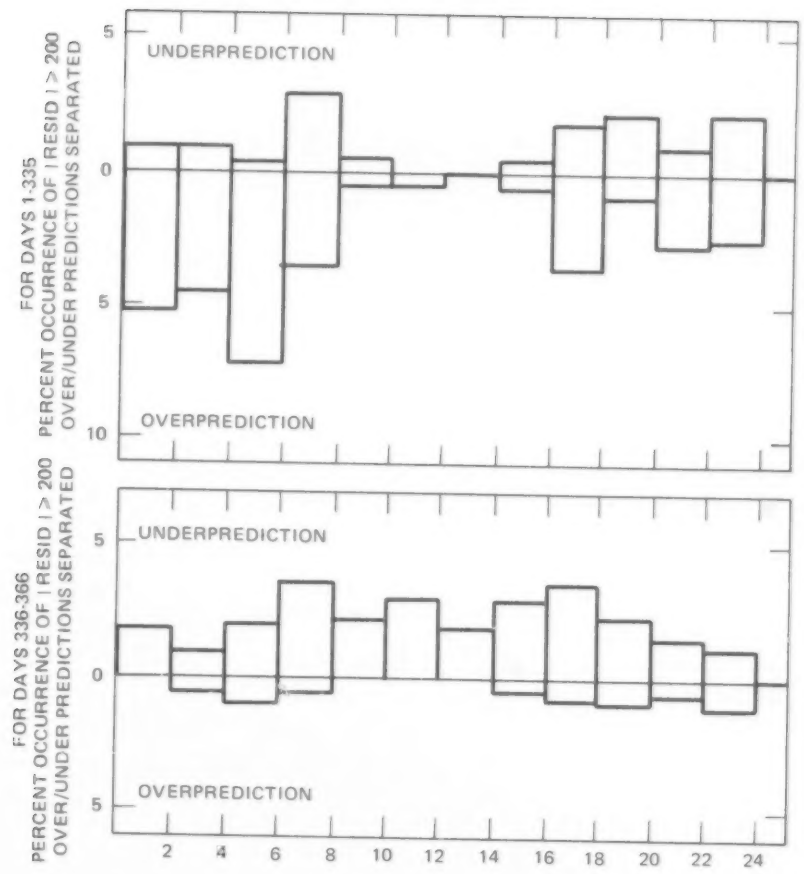


Figure D-5. Distribution of Site 104 residuals by time of day.

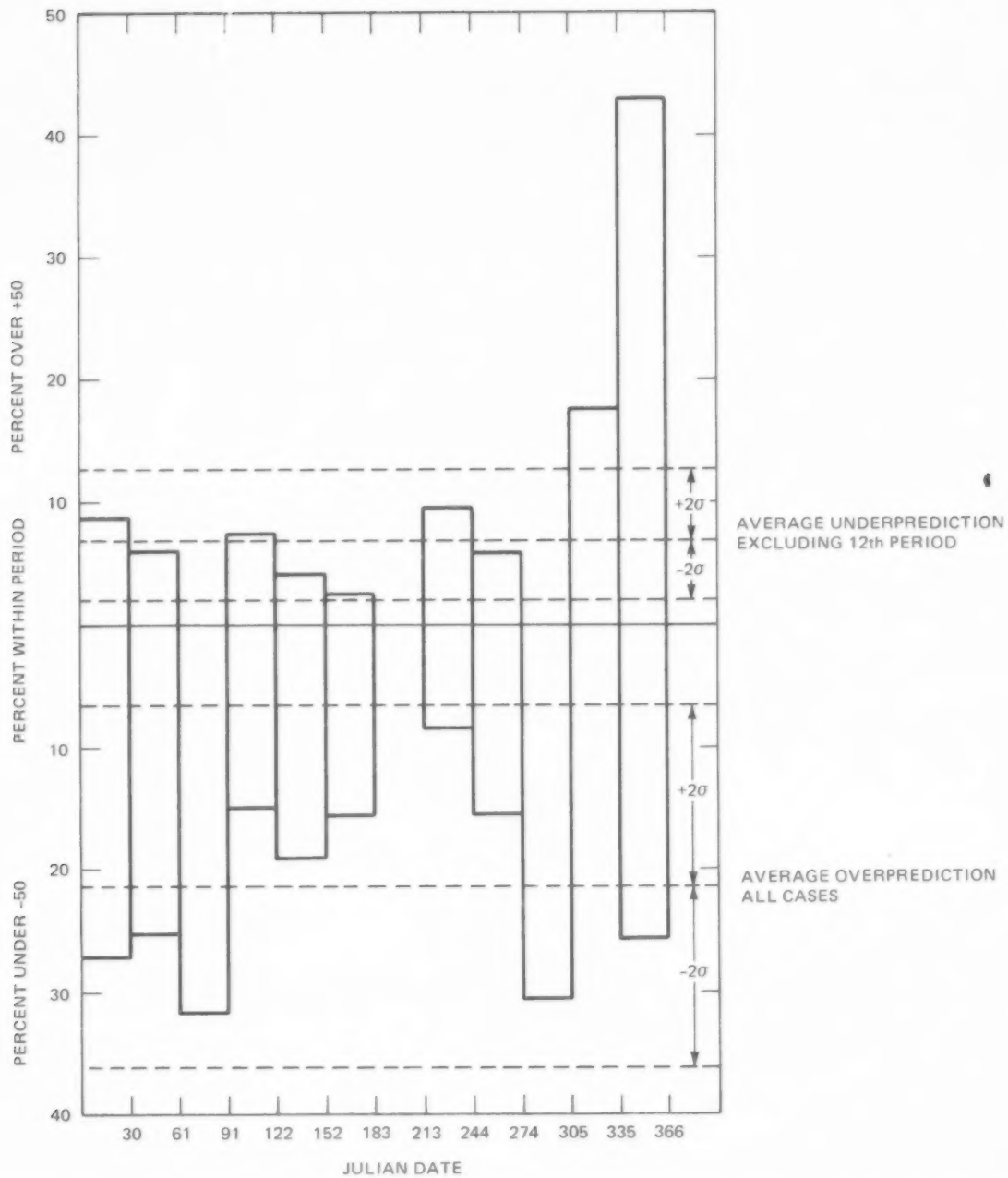


Figure D-6. Distribution of overprediction and underprediction for 1976.